Supplemental Information

Supplementary Tables (All provided as separate excel).

Supplementary Table 1: Sample summary

Summary information for each sample is listed, including sample ID, cohort, tissue type (frozen/FFPE), availability of paired normal, cell of origin (COO) and comprehensive consensus clustering (CCC) subtypes, QC metrics (coverage tumor and normal, TiN, etc.), purity, ploidy, cluster association, summary statistics for EBV and genetic features (total mutation count and density, driver mutation count and density, non-synonymous mutation rate, total and driver SCNAs, number of chromosomal rearrangements).

Supplementary Table 2: Patient characteristics

a, Patient characteristics are summarized, including ID, cohort, age, gender, morphological subtype, IPI and its factors, PFS, OS, R-CHOP-like treatment (yes/no). **b,** Patient characteristics by cohort. The p-values were obtained by a Kruskal-Wallis rank-sum test (age) or a Fisher's exact test (all other categorical variables). All p-values were two-sided with no adjustment for multiple comparisons.

Supplementary Table 3: Significantly mutated genes

a, Mutated genes ranked by their significance values obtained from MutSig2CV. **b,** Genes with significant spatial clustering within a protein structure as detected by *CLUMPS*. **c,** Genes with significant spatial clustering at protein-protein interfaces as detected by *EMPRINT*. **d,** MAF file of all mutations of samples with a paired normal. **a-d,** Analyses have been performed in the full cohort (n=304). **e,** Two-sided Fisher's exact test comparing frequencies of CCGs in tumor-only (TO, n=169) and tumor-normal paired samples (TN, n=135). Ranked by q-value**. f,** Two -sided Fisher's exact test comparing frequencies of CCGs in samples obtained from FFPE (n=136) and fresh-frozen (n=168) samples after removal of the focal copy number gain peak 21q22.3**.**

Supplementary Table 4: Mutational signature analyses

a, Mutational signature activity in 304 samples. **b,** Mutational signature activity including clustering in 303 samples. **c,** Aging signature enrichment by gene (n=12532 genes). **d,** cAID signature enrichment by gene (n=328 genes). **e,** AID2 signature enrichment by gene (n=967 genes). **c-e,** The p-values were obtained using a one-sided binomial test and the p-values were corrected for multiple hypotheses. Genes that are associated with each signature were identified using a *q*-value cutoff of 0.1. **f,** Cosine similarity of mutational signatures discovered in test sets to evaluate germline and FFPE contamination.

Supplementary Table 5: Chromosomal rearrangements

a, Regions of the targeted bait set for structural variants (SV) detection. **b,** Chromosomal rearrangements as reported by the newly developed *Breakpointer* pipeline. For each event, the table summarizes the chromosomal position of the first and second gene, type of rearrangement, support by which detection algorithm, supporting split read and read pair count of the alternate and reference alleles as well as the calculated cancer cell fractions (CCFs). **c,** Matrix of frequent (at least in 2 samples) chromosomal rearrangement by sample. **d,** Chromosomal rearrangements and the reported CCFs involving *MYC*, *BCL2* and *BCL6* in 31 LBCL cell lines. Same format as b.

Supplementary Table 6: Significant SCNAs and correlation to gene expression

a, List of significant arm-level and focal SCNAs (CN gain and CN loss) with FDR <0.1 as identified by GISTIC2.0 in all 304 DLBCLs. **b,** Summary of focal peaks. Wide peak coordinates, genes within wide peaks

and summary statistics of within-peak genes with positive correlation to gene expression are listed. **c,** Detailed list of all genes with significant correlation between focal-peaks and associated cis-acting gene expression (FDR<0.25; FC>1.2). **d,** Summary of arm-level alterations. Genes within arm-level alterations and summary statistics of arm-level genes with positive correlation to gene expression are listed. **e,** Detailed list of all genes with significant correlation (FDR<0.25; FC>1.2) between arm-level alterations and associated cis-acting gene expression. **f,** Summary of focal plus arm-level alterations. Wide peak coordinates of focal-peaks, genes within wide peaks and summary statistics of within-peak genes with positive correlation between arm-level or focal-peaks to gene expression are listed. **g,** Detailed list of all genes with significant correlation (FDR<0.25; FC>1.2) between focal and arm-level alterations and associated cis-acting gene expression. **b-g,** Correlation of genes in GISTIC-defined alterations was performed for all samples with full available gene expression profiles (n=137). **h,** Gene sets used in gene set enrichment analysis in Fig. S13f-h.

Supplemental Table 7: Univariate and multivariate outcome associations of genetic drivers

a, Univariate Cox model for all genetic driver alterations with at least 3% events in the R-CHOP treated cohort (n=259) for PFS and OS. Ranked by significance (q-value). **b,** Cox regression models of IPI with all significant factors from the univariate analyses for PFS (n=254) and OS (n=259) in the R-CHOP treated cohort.

Supplementary Table 8: Gene sample matrix and features of consensus clusters

a, Gene sample matrix. For each of the 304 samples, each of the 159 genetic "drivers" with a frequency ≥3% are listed. Mutations (0, absent; 1, synonymous; 2, non-synonymous); SCNAs (no SCNA, 0; low grade SCNA,1; high grade SCNA, 2); Chromosomal Rearrangements (SV; absent, 0; present, 3). **b,** Consensus clustering results. Cophenetic coefficient for k=4 to k=10 clusters, membership of each sample and silhouette values for "Best cluster" (k=5); **c,** Feature selection for each cluster, C1-C5 (n=292), by one-sided Fisher's exact test.

Supplementary Table 9: Clinical features, features across clusters and gene sets tested for an enrichment

a, Summary table of clinical features by cluster. All p-values were obtained by a two-sided Fisher's Exact test; p-values were not corrected for multiple comparisons. **b,** Pairwise comparisons of significant results from a. The p-values are obtained by a two-sided Kruskal-Wallis rank-sum test (purity and ploidy) or two sided Fisher's Exact test. p-values were not corrected for multiple comparisons.

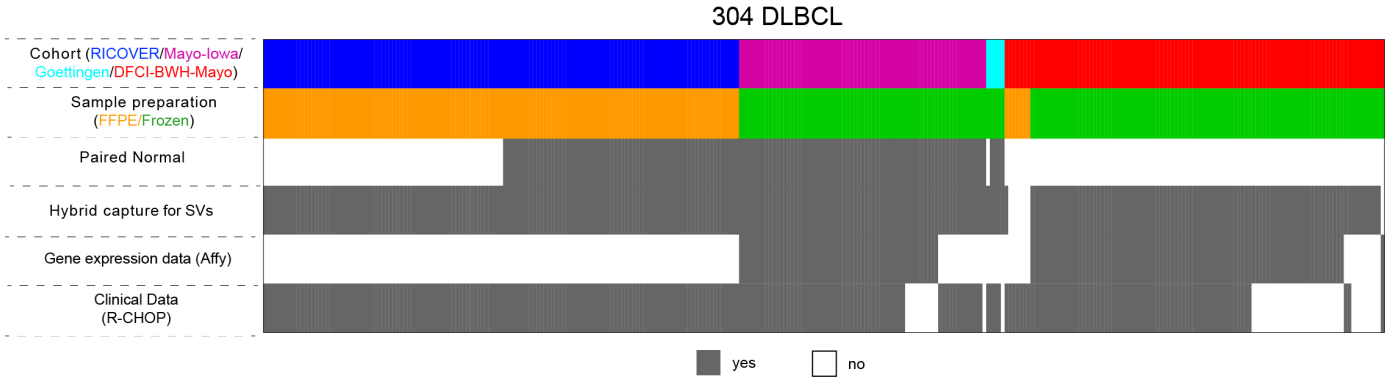
Supplementary Table 10: Ordering analyses

a, CCF-Matrix for all 158 driver alterations. **b,** Occurring and modelling of clonal-subclonal pairs by cluster**.** **c,** Results of ordering analyses for clonal-subclonal pairs powered to achieve a q-value<0.1. **b-c,** Ordering analysis was done for clusters C1-C5 (n=292). The p-values were obtained by a two-sided binomial test. and were corrected using the method of Benjamini and Hochberg.

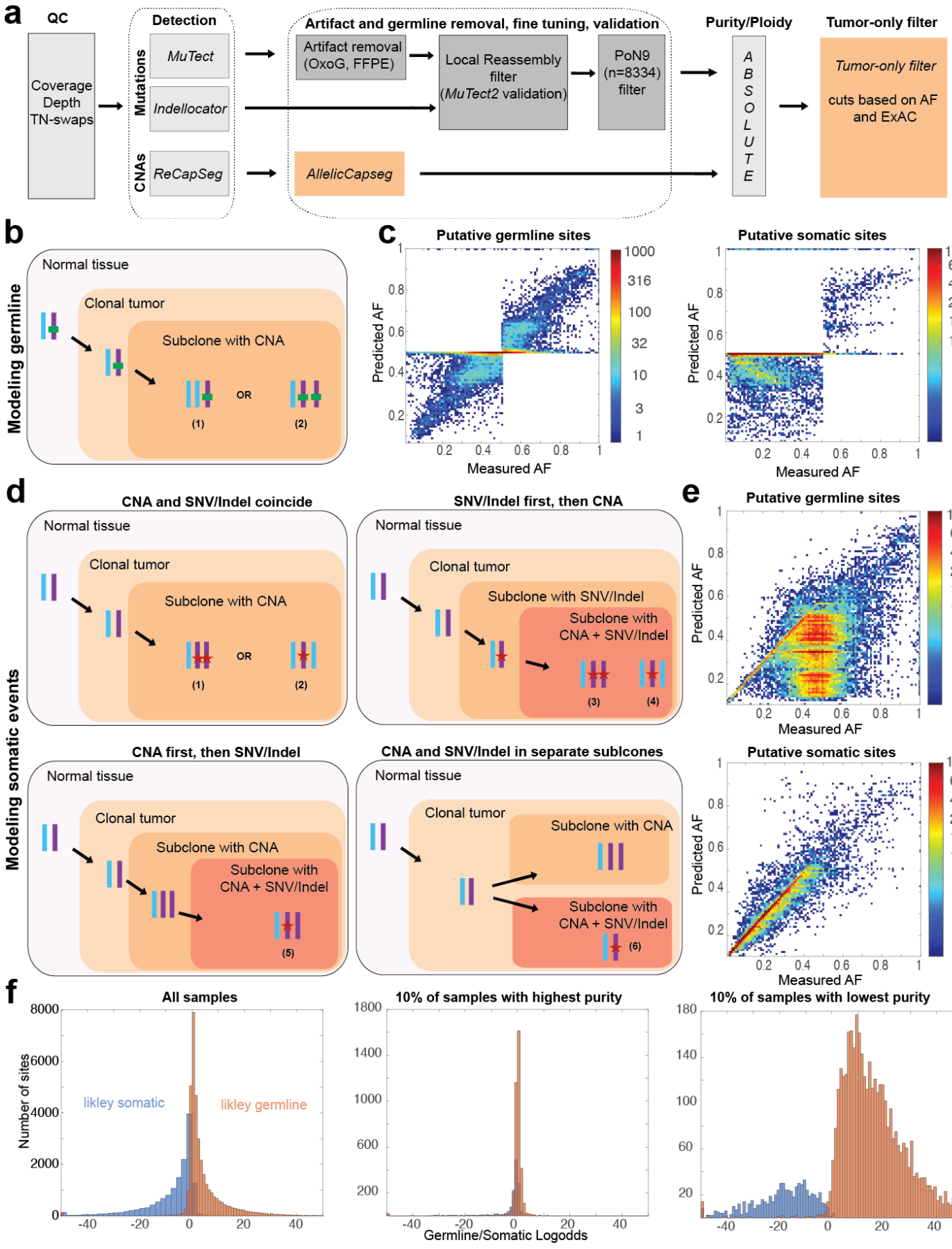
Supplementary Table 11: **Outcome analyses of clusters**

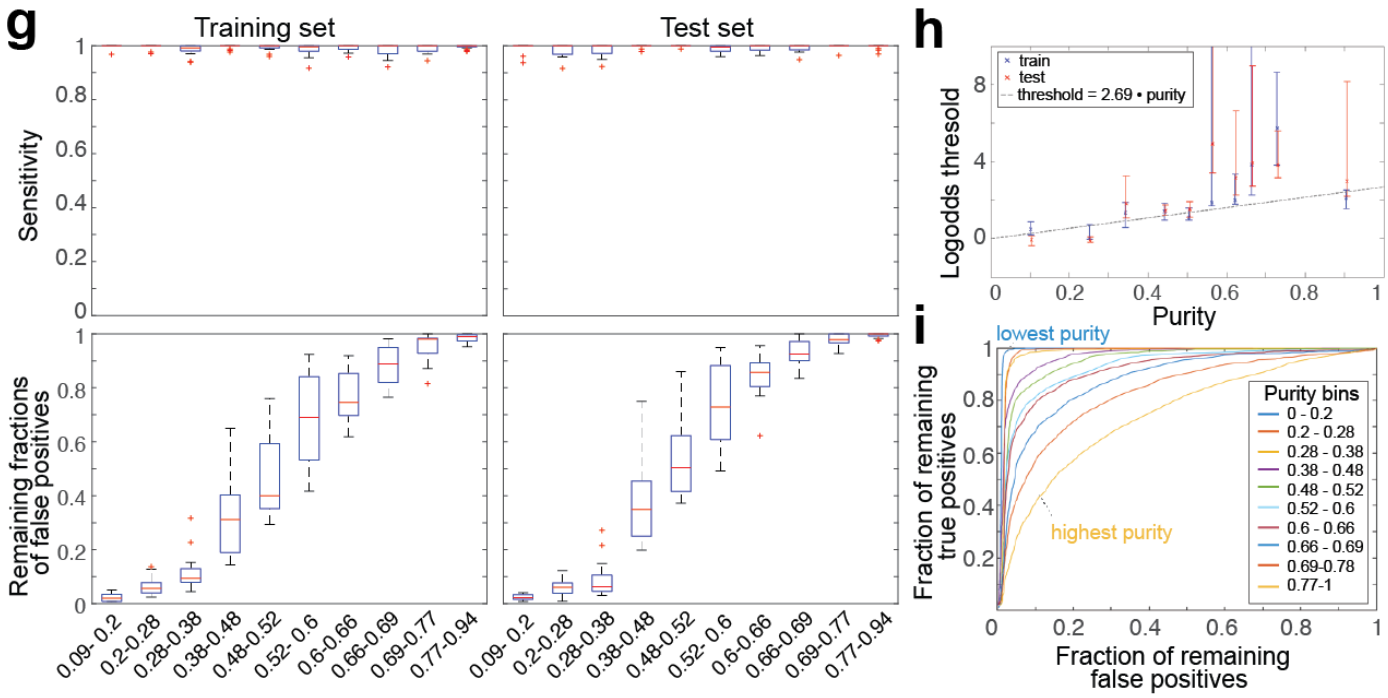
a, PFS and OS survival proportions and 95% confidence intervals every 12 months for each cluster in the R-CHOP treated cohort (n=259). **b,** Multivariate model of clusters and IPI (with comparison to IPI-only model) for PFS and OS in the R-CHOP treated cohort (n=259).

Supplementary Figures



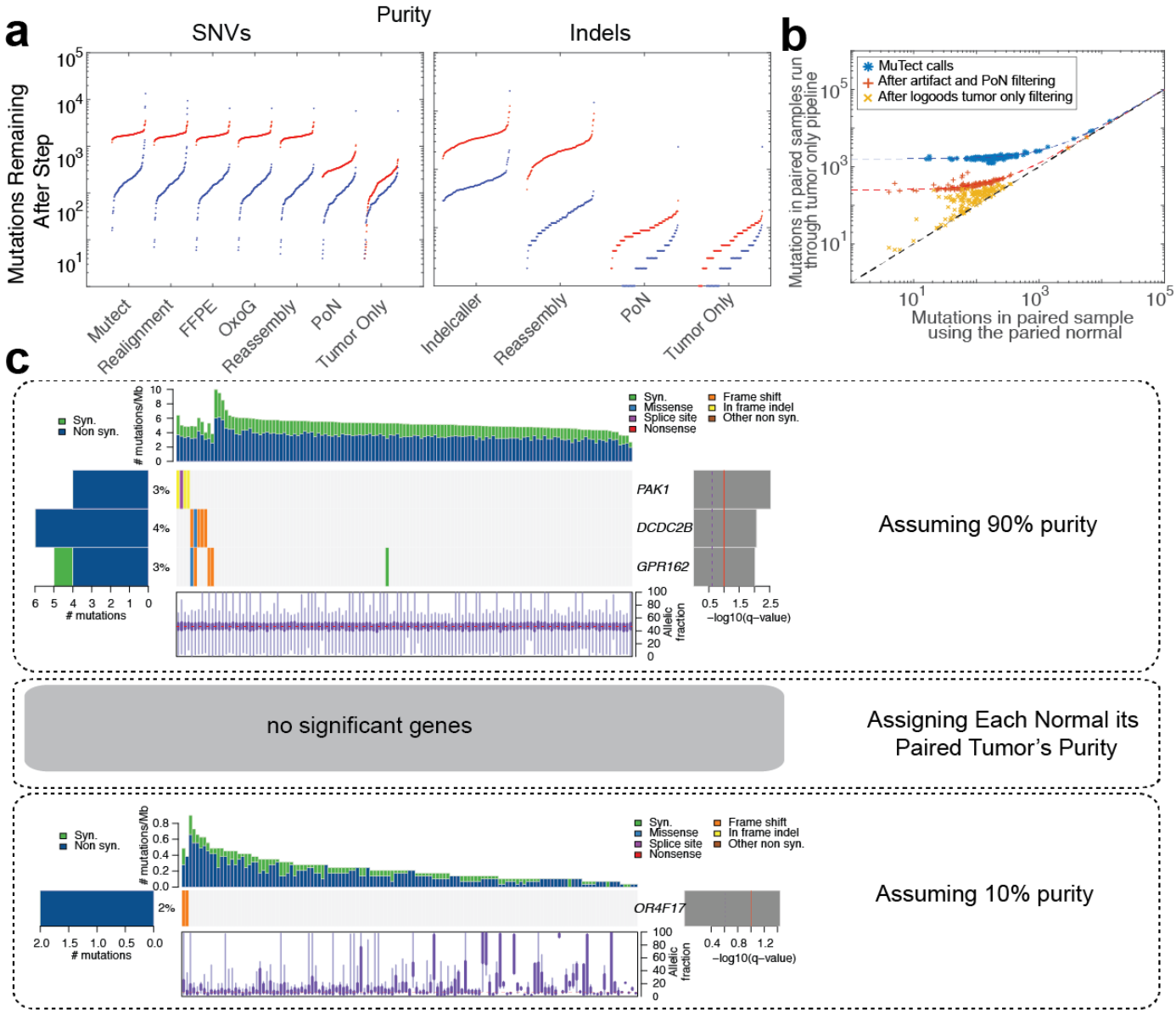
Supplementary Figure 1. Composition of the dataset. The dataset includes 304 newly diagnosed DLBCLs from 4 cohorts (129 samples from the RICOVER60 trial[1]; 67 samples from a Mayo/Iowa cohort, of which 51 WES were previously reported[2,3]; 5 samples from the University of Göttingen, Germany; 103 samples from a DFCI/BWH/Mayo cohort[4]; top row) including DNA-derived from formalin-fixed paraffin embedded (FFPE) or frozen tissue (second row). DLBCLs with paired normal samples are indicated (third row). Samples used for targeted sequencing analyses of recurrent structural variants (SVs) (fourth row) and transcriptional profiling (fifth row) are noted. DLBCLs from patients who were treated with state-of-the-art therapy (R-CHOP) and have long-term follow-up are also indicated below (bottom line).

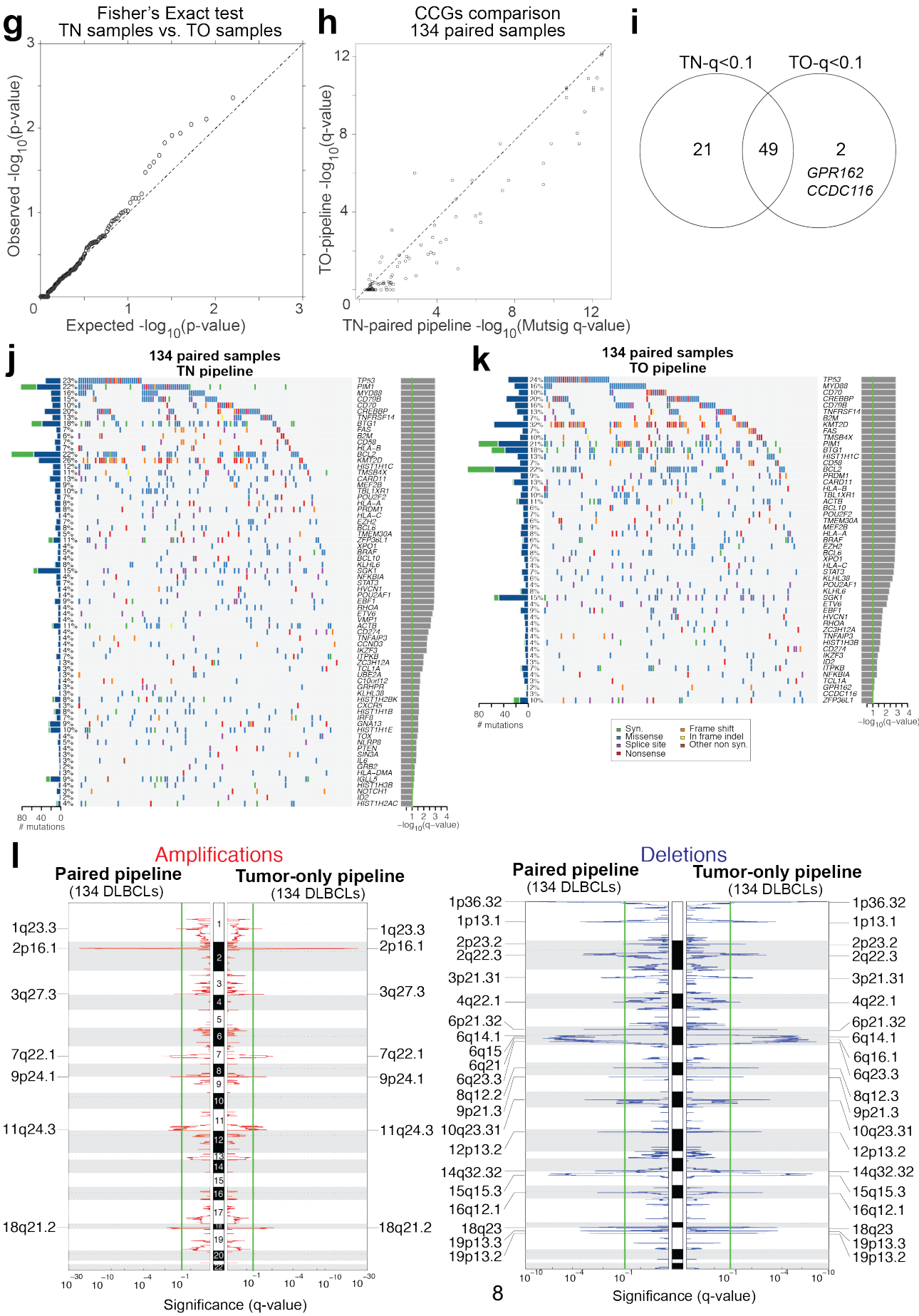

a
QC
Coverage
Depth
TN-swaps
Detection
Mutations
MuTect
Indellocator
CNAs
ReCapSeg
Artifact and germline removal, fine tuning, validation
Artifact removal (OxoG, FFPE)
Local Reassembly filter (MuTect2 validation)
PoN9 (n=8334) filter
AllelicCapseg
Purity/Ploidy
ABSOLUTE
Tumor-only filter
Tumor-only filter
cuts based on AF and ExAC
b
Modeling germline
Normal tissue
Clonal tumor
Subclone with CNA
OR
(1)
(2)
c
Putative germline sites
Putative somatic sites
Predicted AF
Measured AF
d
Modeling somatic events
CNA and SNV/Indel coincide
SNV/Indel first, then CNA
CNA first, then SNV/Indel
CNA and SNV/Indel in separate sublcones
Subclone with SNV/Indel
Subclone with CNA + SNV/Indel
(3)
(4)
(5)
(6)
e
Putative germline sites
Putative somatic sites
f
All samples
10% of samples with highest purity
10% of samples with lowest purity
Number of sites
likley somatic
likley germline
Germline/Somatic Logodds

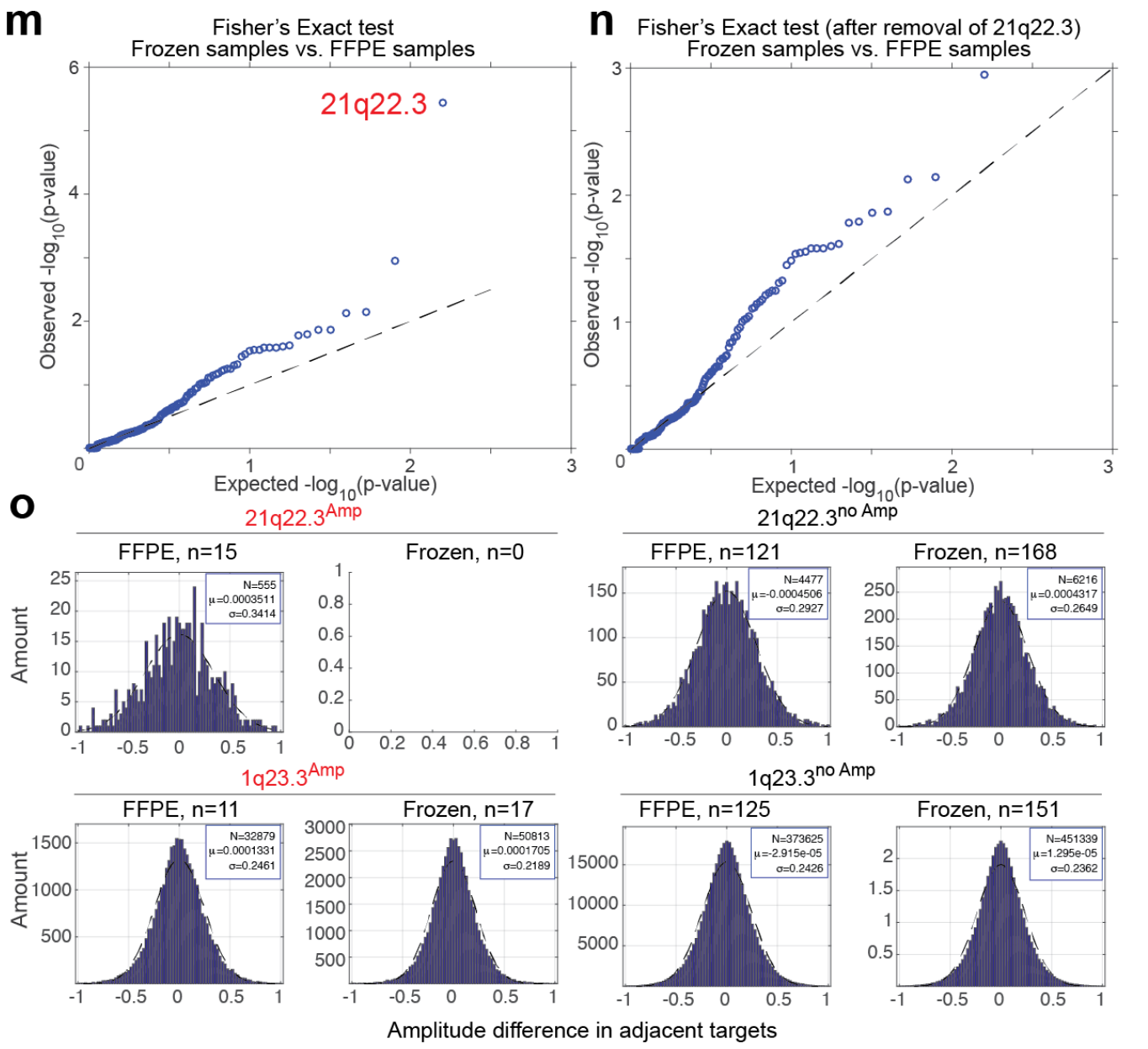


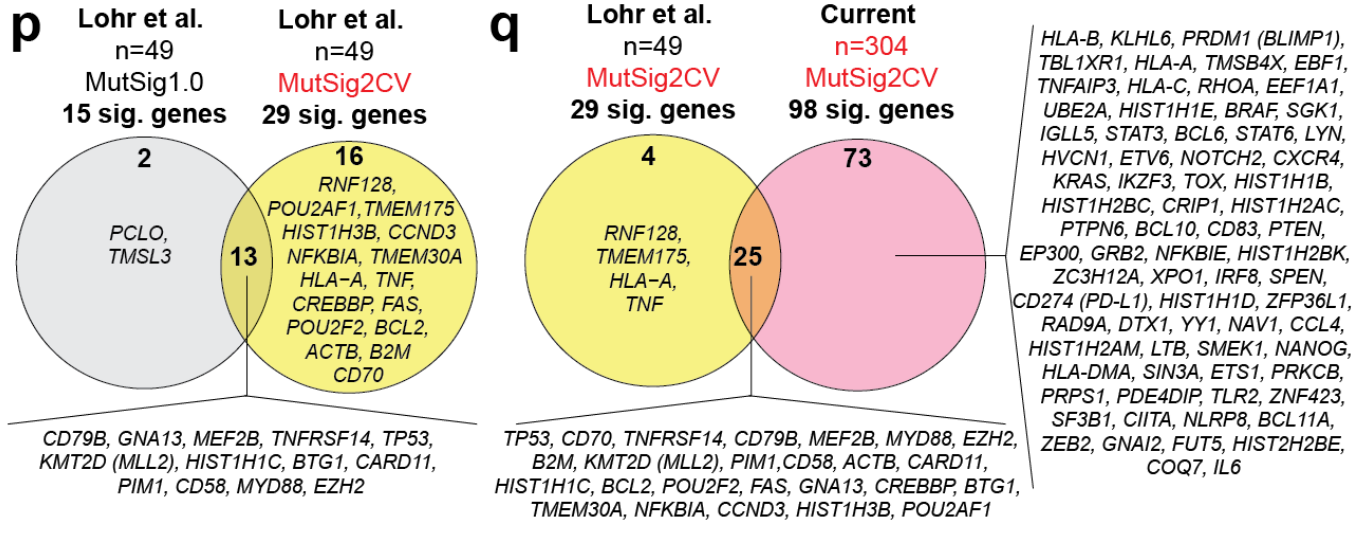
Supplementary Figure 2. Mutation and SCNA pipeline and concepts for the germline somatic log odds filter for tumor-only samples. a, Analysis pipeline used for mutation and SCNA detection in the DLBCL cohort. Steps highlighted in red were run with slight adjustments for tumor-only samples, namely the tumor was used for het site detection in *AllelicCapseg* and the additional filter that calculates the log odds ratio of an event being germline or somatic based on purity and CCF was applied. **b,** Cartoon demonstrating the conditions that would affect a germline het site's allele fraction (AF). In the normal component of the sample, a germline het site should have a 50% variant AF and should only deviate from this if there is a copy number alteration that affects it. **c,** Scatter plots demonstrate the correlation between the actual AF and the predicted AF using the model assuming the event is germline. In the left plot, the events are limited to those that were detected in the analysis where the paired samples were run without their paired normal, and are therefore putative germline events, while the plot on the right represents sites that were detected when paired samples were run with the paired normal, and are therefore putative somatic mutations. When using the model that assumes a site is germline, there are many events showing a positive correlation between the predicted AF and observed AF for putative germline events, while the same correlation is not observed in the putative somatic events. **d,** Cartoon demonstrating the conditions that would affect a somatic events' allele fraction, which should be 0% in the normal component of the sample as well as any part of the tumor that is not part of the subclone where the event developed. The allele fraction can also be shifted by any copy number event that affects the region of the somatic event. All scenarios are considered in modeling the predicted AF for somatic events, and the most likely one is used to compare against the germline prediction. **e,** Scatter plots are the same as in c, except the predicted and actual AFs for the somatic model are plotted. There is a much stronger correlation for the actual AF and predicted AF in the putative somatic events than putative germline events. **f,** Once the most likely AF is calculated, the likelihood that it is consistent with the observed AF is calculated for the germline model and somatic model, and the log odds ratio is plotted here for putative germline and somatic events. The log odds diverge more for low-purity samples because the AF for somatic events changes more as a result of the fact that they have 0% AF in the normal component of the tumor while germline events are present at the 50% rate. **g-i,** To determine the sensitivity and false discovery rate of the tumor-only pipeline, each of the paired samples run as tumor-only

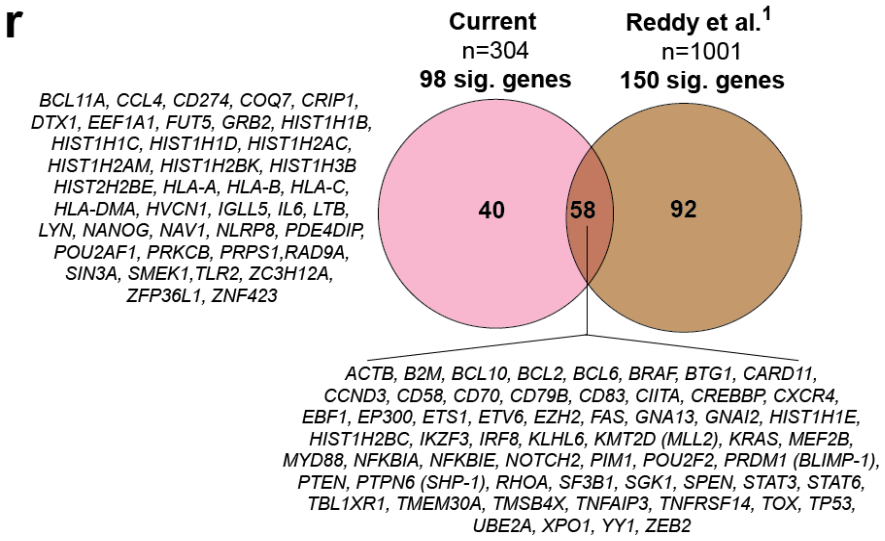
(n=147) were randomly split assigned into a training set and a test set such that half of the data was in each set (n_1=74 and n_2=73). Each set was then divided by purity so that 10% of the data was placed in each bin, and the cutoff that yielded 99% sensitivity (retained 99% of putative somatic mutations) was calculated for each purity bin. These cutoffs were then used to fit a formula based on purity to determine the cutoff for each sample. The resulting cutoff was applied to the other dataset. **g,** Distribution of sensitivity and false positive rate for patients within each purity decile as boxplot (median, red line; inter-quartile range, box; approximately 2.7 standard deviations of median, whiskers). The false discovery rate is notably very low for low purity samples due to the allele shift in somatic mutations down from 50%, while very high for more pure samples but sensitivity does not go below 90% for any purity bin. **h,** Logodds thresholds for 99% sensitivity calculated within each decile for each training/test set (n_1=74 and n_2=73) that was used to calculate the logodds thresholds for the other set. The points represent the mean logodds threshold in each decile that will yield a true positive rate of 99%, while the error bars represent one standard deviation assuming a beta distribution. The plotted line represents a linear fit to the training set. **i,** ROC curve of different log odds cutoffs at different purities.

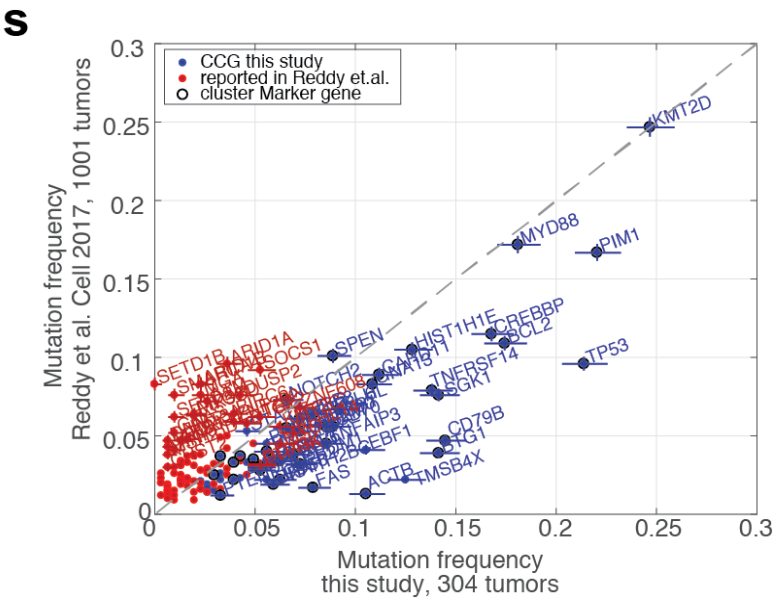

a
Purity
SNVs
Indels
Mutations Remaining After Step
Mutect
Realignment
FFPE
OxoG
Reassembly
PoN
Tumor Only
Indelcaller
Reassembly
PoN
Tumor Only
b
MuTect calls
After artifact and PoN filtering
After logoods tumor only filtering
Mutations in paired samples run through tumor only pipeline
Mutations in paired sample using the paried normal
c
Syn.
Non syn.
mutations/Mb
Frame shift
In frame indel
Other non syn.
Missense
Splice site
Nonsense
PAK1
DCDC2B
GPR162
mutations
Allelic fraction
−log10(q−value)
Assuming 90% purity
no significant genes
Assigning Each Normal its Paired Tumor's Purity
OR4F17
Assuming 10% purity
d
AllelicCapseg segment amplitude
r2=0.9959
Measured using tumor-only pipeline
Measured using paired pipeline
e
ABSOLUTE-defined ploidy
r2=0.9759
Estimated using tumor-only pipeline
Estimate using paired pipeline
f
ABSOLUTE-defined purity
r2=0.8795
Estimated using tumor-only pipeline
Estimate using paired pipeline

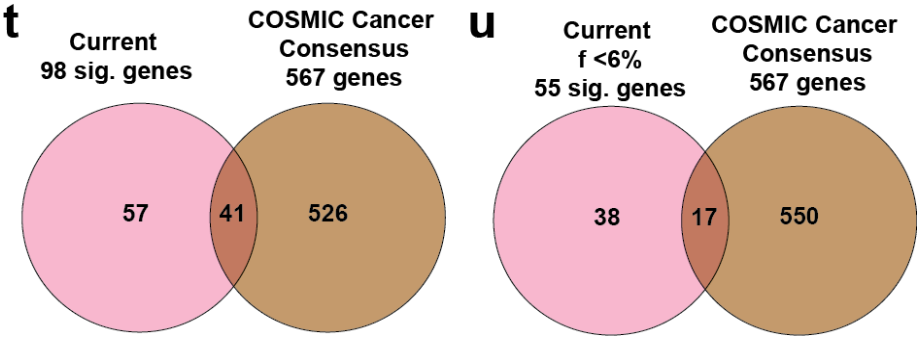

g
Fisher's Exact test
TN samples vs. TO samples
Observed -log10(p-value)
Expected -log10(p-value)
h
CCGs comparison
134 paired samples
TO-pipeline -log10(q-value)
TN-paired pipeline -log10(Mutsig q-value)
i
TN-q<0.1
TO-q<0.1
21
49
2
GPR162
CCDC116
j
134 paired samples
TN pipeline
mutations
−log10(q-value)
k
134 paired samples
TO pipeline
mutations
−log10(q-value)
Syn.
Missense
Splice site
Nonsense
Frame shift
In frame indel
Other non syn.
l
Amplifications
Paired pipeline
(134 DLBCLs)
Tumor-only pipeline
(134 DLBCLs)
1q23.3
2p16.1
3q27.3
7q22.1
9p24.1
11q24.3
18q21.2
Significance (q-value)
Deletions
Paired pipeline
(134 DLBCLs)
Tumor-only pipeline
(134 DLBCLs)
1p36.32
1p13.1
2p23.2
2q22.3
3p21.31
4q22.1
6p21.32
6q14.1
6q15
6q21
6q23.3
8q12.2
9p21.3
10q23.31
12p13.2
14q32.32
15q15.3
16q12.1
18q23
19p13.3
19p13.2
6q16.1
8q12.3
9p21.3
Significance (q-value)


m
Fisher's Exact test
Frozen samples vs. FFPE samples
21q22.3
Observed -log10(p-value)
Expected -log10(p-value)
n
Fisher's Exact test (after removal of 21q22.3)
Frozen samples vs. FFPE samples
Observed -log10(p-value)
Expected -log10(p-value)
o
21q22.3Amp
FFPE, n=15
Frozen, n=0
21q22.3no Amp
FFPE, n=121
Frozen, n=168
1q23.3Amp
FFPE, n=11
Frozen, n=17
1q23.3no Amp
FFPE, n=125
Frozen, n=151
Amount
Amplitude difference in adjacent targets


p
Lohr et al.
n=49
MutSig1.0
15 sig. genes
Lohr et al.
n=49
MutSig2CV
29 sig. genes
2
PCLO, TMSL3
13
16
RNF128, POU2AF1,TMEM175 HIST1H3B, CCND3 NFKBIA, TMEM30A HLA−A, TNF, CREBBP, FAS, POU2F2, BCL2, ACTB, B2M CD70
CD79B, GNA13, MEF2B, TNFRSF14, TP53, KMT2D (MLL2), HIST1H1C, BTG1, CARD11, PIM1, CD58, MYD88, EZH2
q
Lohr et al.
n=49
MutSig2CV
29 sig. genes
Current
n=304
MutSig2CV
98 sig. genes
4
RNF128, TMEM175, HLA−A, TNF
25
73
TP53, CD70, TNFRSF14, CD79B, MEF2B, MYD88, EZH2, B2M, KMT2D (MLL2), PIM1,CD58, ACTB, CARD11, HIST1H1C, BCL2, POU2F2, FAS, GNA13, CREBBP, BTG1, TMEM30A, NFKBIA, CCND3, HIST1H3B, POU2AF1
HLA-B, KLHL6, PRDM1 (BLIMP1), TBL1XR1, HLA-A, TMSB4X, EBF1, TNFAIP3, HLA-C, RHOA, EEF1A1, UBE2A, HIST1H1E, BRAF, SGK1, IGLL5, STAT3, BCL6, STAT6, LYN, HVCN1, ETV6, NOTCH2, CXCR4, KRAS, IKZF3, TOX, HIST1H1B, HIST1H2BC, CRIP1, HIST1H2AC, PTPN6, BCL10, CD83, PTEN, EP300, GRB2, NFKBIE, HIST1H2BK, ZC3H12A, XPO1, IRF8, SPEN, CD274 (PD-L1), HIST1H1D, ZFP36L1, RAD9A, DTX1, YY1, NAV1, CCL4, HIST1H2AM, LTB, SMEK1, NANOG, HLA-DMA, SIN3A, ETS1, PRKCB, PRPS1, PDE4DIP, TLR2, ZNF423, SF3B1, CIITA, NLRP8, BCL11A, ZEB2, GNAI2, FUT5, HIST2H2BE, COQ7, IL6







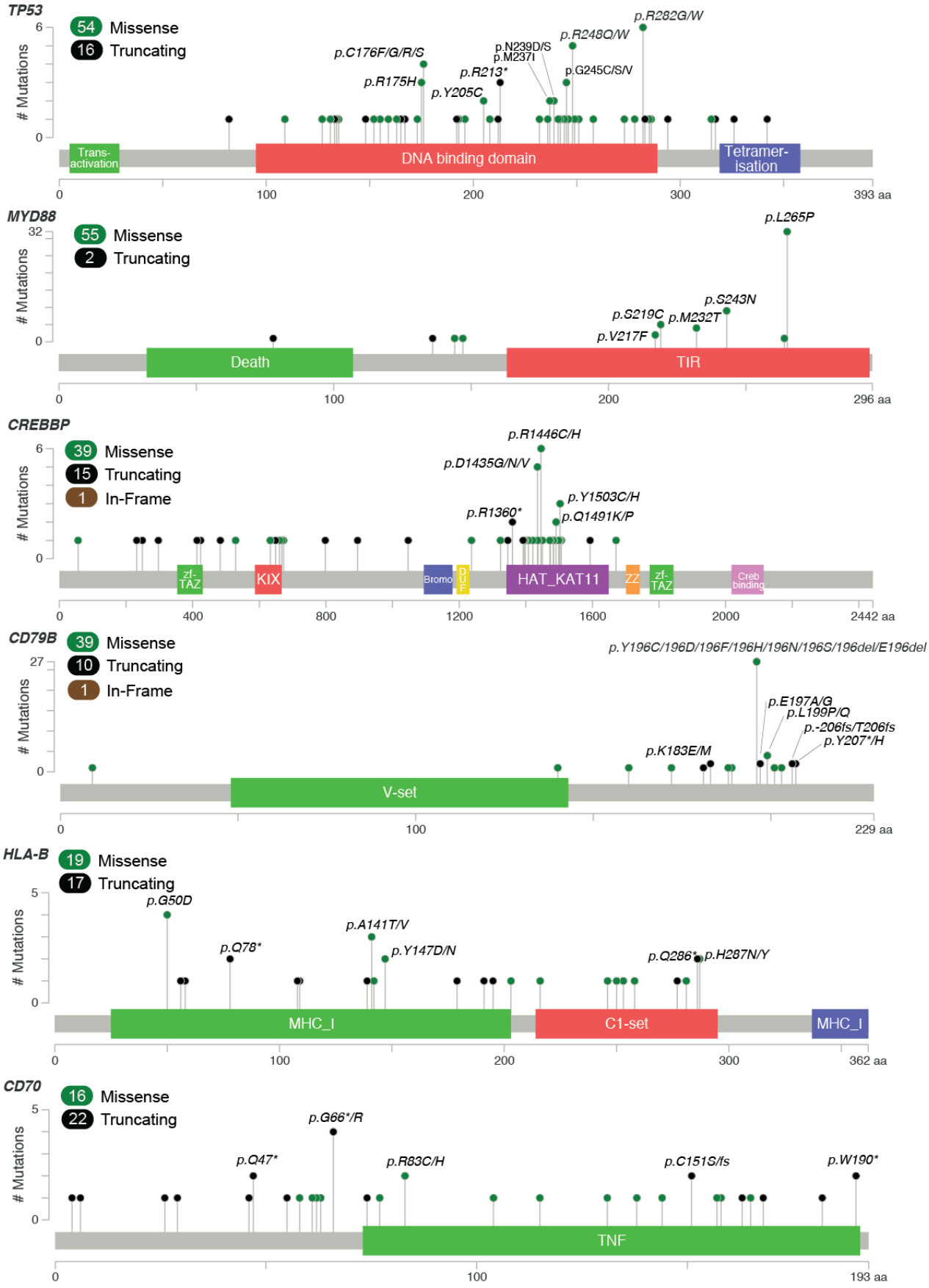
Supplementary Figure 3. Tumor-only filter performance. a, Distribution of total event counts for SNVs and Indels after each step of the analysis for tumor-only (red) and paired (blue) samples. **b,** Mutation counts for the paired samples when run through the paired analysis (x) and tumor-only pipeline (y). After the full tumor-only pipeline, many samples have approximately as many events called as the paired analysis, but none have fewer, which is consistent with the high sensitivity and low false positive rate demonstrated in figure a. **c,** The paired normals of our DLBCL cohort were run as "tumors" through the tumor-only pipeline in order to show that after the tumor-only pipeline there are few if any recurrent events that could be contaminating the list of cancer consensus genes. Since these samples should not have a tumor component,

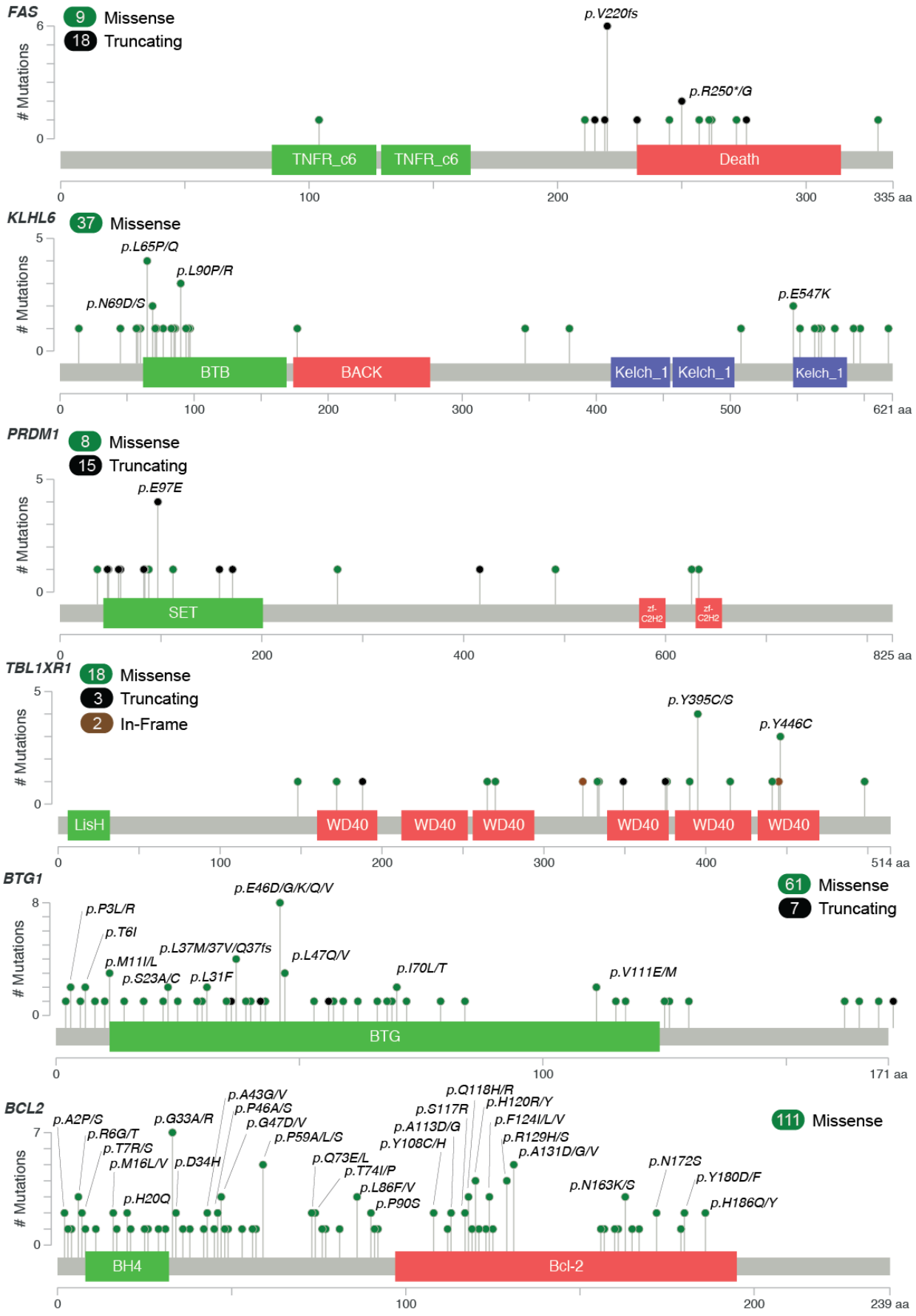
trials were tried where each sample was assigned 90% purity, the same purity as its paired tumor, and 10%, all of which yield up to 3 significant genes, which is a very small number given the sample size. **d-f,** Comparison of *AllelicCapseg*-defined amplitudes at points selected throughout the genome every 1 megabase (d), *ABSOLUTE*-defined ploidy (e) and purity (f) using 147 paired-samples processed either through the tumor-only pipeline or the paired pipeline. r^2 values are the Spearman's correlation coefficients squared**. g,** A two-sided Fisher's exact test was applied to each of the 159 putative driver events to determine if there was a significant bias in samples that were tumor-only (TO) vs. paired (TN) samples. The graph shows that each of the observed p-values is as it would be randomly expected, suggesting that there is no bias of the tumor-only analyses towards any of the 159 putative drivers. **h,** Separate *Mutsig2CV* analyses were performed for the 134 paired samples using the patient-matched normal sample or applying our tumor-only pipeline. Scatter plot of *Mutsig2CV* q-values using (x-axis) or not using (y-axis) the patient-matched normal sample (q-values <0.5 are plotted). **i,** Venn-diagram of significant mutated genes as identified by *Mutsig2CV* (q-value <0.1) using the patient-matched normal (TN, left, n=135) or the tumor-only (TO, right, n=169) pipeline. **j,k,** CoMut plot of CCGs detected in 134 samples with available paired normal using the patient-matched normal sample (j) or the tumor-only pipeline (k). **l,** *GISTIC2.0*-defined recurrent SCNAs in 134 paired samples were analyzed within the CN-pipeline using either the patient-matched normal (mirror plot, left side) or an unrelated normal (mirror plot, right side). Separate mirror plots for amplifications (left) and deletions (right). **m-o,** A two-sided Fisher's exact t-test was performed for all 159 genetic events and detected a significant bias of the focal 21q22.3 copy number gain in the Q-Q plot (m). The 21q22.3 focal gain was found to be significantly overrepresented in FFPE samples (q = 0.00058) with all 15 detected events occurring in FFPE samples (n=136) compared to 0 events found in the frozen samples (n=168) (m). This significant test combined with follow up review of the noise of targets in the region (o), led to removal of 21q22.3 copy gain from the analysis as a probable artifact. Q-Q plot for the two-sided Fisher's exact t-test after removal of 21q22.3 reveals no significant enrichment (q<0.1) for any of the remaining 158 genetic drivers (n). **o,** As measurement of CN noise we plotted the distribution of amplitude difference in adjacent targets of the focal amplification peak, 21q22.3 (top row), and a representative second focal amplification peak, 1q23.3 (bottom row) and compared the SD (σ) of this common noise metric[5]. For each alteration, we visualized the noise metric in FFPE and frozen samples and for each group separated cases that harbor or lack the event. The number of FFPE or frozen samples with the indicated alterations, n, is shown. N in each histogram is the number of total segments/probe measurement visualized. Notably, focal gain of 21q22.3 was only found in FFPE samples and had the highest σ of all focal events, which prompted the removal from all analyses. **p,** Comparison of significantly mutated genes prioritized by *MutSig1.0* vs. *MutSig2CV*[6] algorithms in an equal sized DLBCL data set (previously published Lohr et al dataset[2]) of 49 DLBCL samples. **q,** Comparison of significantly mutated genes prioritized by *MutSig2CV*[6] algorithm in previously published Lohr dataset and the current data set. Of note, the Lohr set is part of the current 304 tumor dataset. **r,** Comparison of significantly mutated genes prioritized by *MutSig2CV*[6] algorithm in this dataset to those identified with *MutSigCV* in Reddy et al.[7] **s,** Scatter plot comparing the frequencies of CCGs reported by Reddy et al. in our series (x-axis) and the Reddy et al. paper (y-axis). Blue genes are found in both studies and red genes are only found in Reddy et al. Of note, the frequency of the 40 genes exclusively identified in our series (Fig. S3r, left) could not be assessed in Reddy et al. A circle around the gene indicate landmark genes for clusters. Error bars represent 1-σ confidence intervals. Importantly, the frequency of certain well known DLBCL mutations, such as *TP53, CREBBP* or *CD79B,* is significantly lower in Reddy et al. **t,** Overlap of 98 CCG genes (q<0.1) to COSMIC Cancer Gene consensus (genes causally implicated in cancer, Nat Rev. Article, https://cancer.sanger.ac.uk/census). **u,** Overlap of 55 low frequency CCG genes (q<0.1, f<6%) to COSMIC Cancer Gene consensus genes.

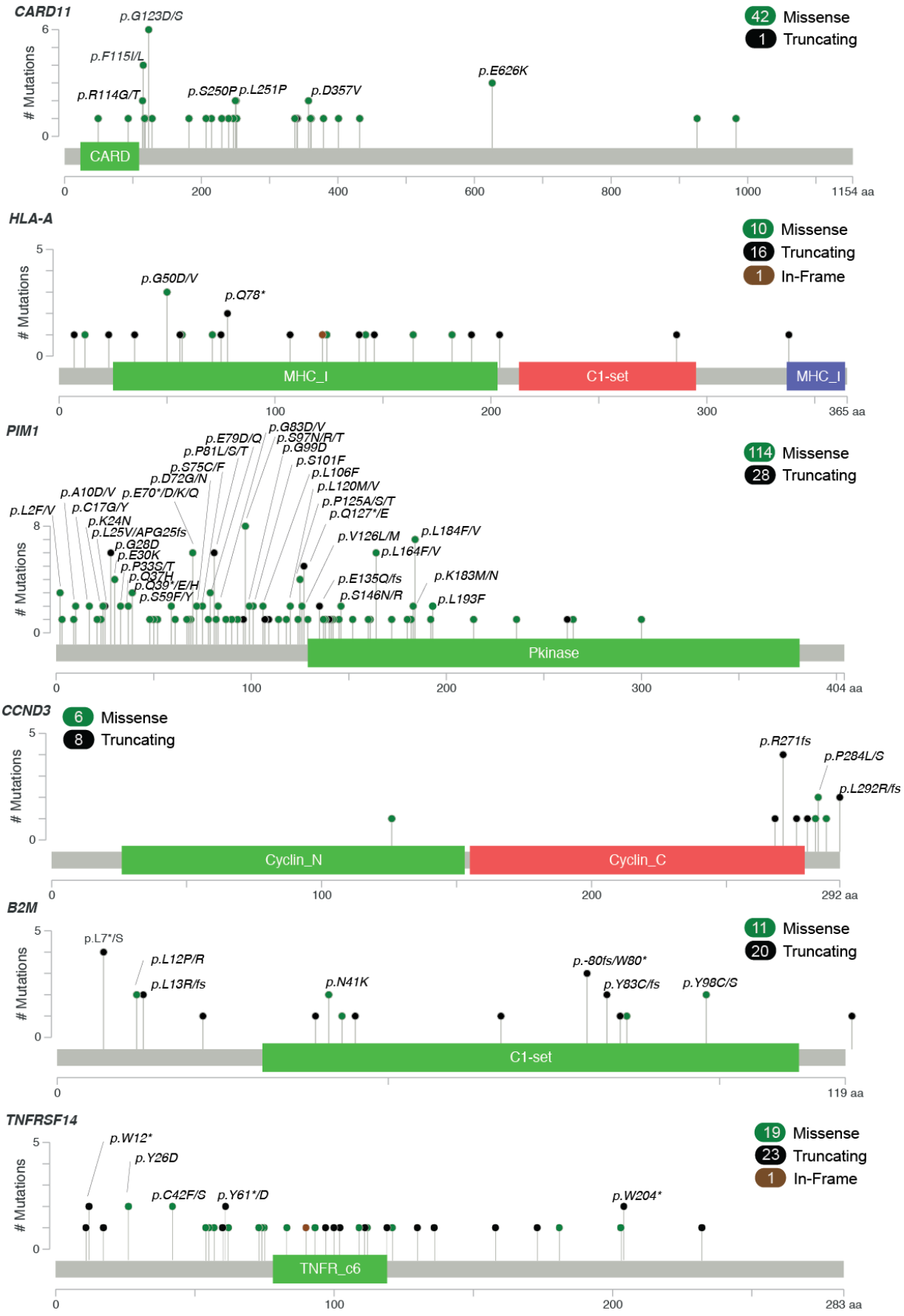
Index for mutation diagrams (lollipop figures)

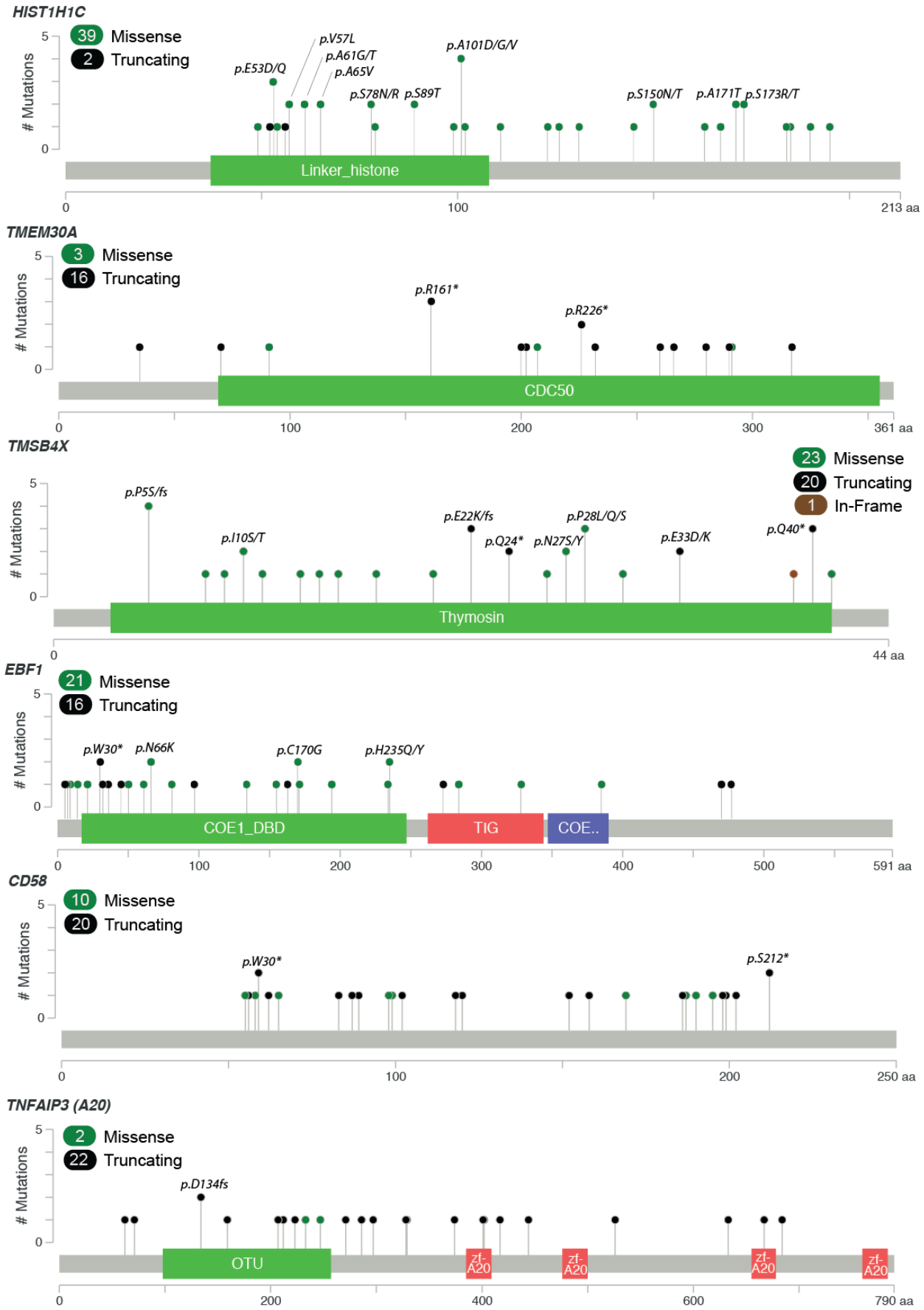
Gene	Longname	nnon	npat	nsite	q	Stickfig. at page
ACTB	actin, beta	37	32	27	7.90E-08	20
B2M	beta-2-microglobulin	31	26	22	1.54E-12	16
BCL10	B-cell CLL/lymphoma 10	21	16	16	3.16E-05	23
BCL11A	B-cell CLL/lymphoma 11A (zinc finger protein)	11	10	10	8.00E-02	29
BCL2	B-cell CLL/lymphoma 2	111	53	76	8.41E-13	15
BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	18	17	16	6.59E-08	20
BRAF	v-raf murine sarcoma viral oncogene homolog B1	19	18	13	1.24E-08	19
BTG1	B-cell translocation gene 1, anti-proliferative	68	43	55	3.67E-13	15
CARD11	caspase recruitment domain family, member 11	43	34	31	1.15E-12	16
CCL4	chemokine (C-C motif) ligand 4	5	4	4	3.78E-03	26
CCND3	cyclin D3	14	14	11	1.39E-12	16
CD274	CD274 molecule	10	8	10	7.89E-04	25
CD58	CD58 molecule	30	19	29	1.16E-11	17
CD70	CD70 molecule	38	27	33	2.02E-13	14
CD79B	CD79b molecule, immunoglobulin-associated beta	50	44	23	2.02E-13	14
CD83	CD83 molecule	24	18	18	3.23E-05	23
CIITA	class II, major histocompatibility complex, transactivator	11	9	10	5.49E-02	28
COQ7	coenzyme Q7 homolog, ubiquinone (yeast)	4	4	3	9.50E-02	30
CREBBP	CREB binding protein (Rubinstein-Taybi syndrome)	64	51	54	2.02E-13	14
CRIP1	cysteine-rich protein 1 (intestinal)	7	7	7	1.42E-05	22
CXCR4	chemokine (C-X-C motif) receptor 4	9	8	8	2.87E-06	21
DTX1	deltex homolog 1 (Drosophila)	48	37	37	2.29E-03	25
EBF1	early B-cell factor 1	37	32	34	1.09E-11	17
EEF1A1	eukaryotic translation elongation factor 1 alpha 1	19	17	16	2.22E-10	18
EP300	E1A binding protein p300	26	25	22	5.34E-05	23
ETS1	v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)	17	15	14	3.09E-02	27
ETV6	ets variant gene 6 (TEL oncogene)	24	21	13	9.48E-07	21
EZH2	enhancer of zeste homolog 2 (Drosophila)	24	22	10	3.02E-10	19
FAS	Fas (TNF receptor superfamily, member 6)	27	24	18	2.02E-13	15
FUT5	fucosyltransferase 5 (alpha (1,3) fucosyltransferase)	4	4	4	8.29E-02	29
GNA13	guanine nucleotide binding protein (G protein), alpha 13	48	33	45	6.47E-08	20
GNAI2	guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 2	11	10	11	8.14E-02	29
GRB2	growth factor receptor-bound protein 2	10	8	9	1.04E-04	24
HIST1H1B	histone cluster 1, H1b	30	26	23	7.17E-06	22
HIST1H1C	histone cluster 1, H1c	41	37	31	2.22E-12	17
HIST1H1D	histone cluster 1, H1d	25	20	22	8.34E-04	25
HIST1H1E	histone cluster 1, H1e	58	39	43	7.88E-09	19
HIST1H2AC	histone cluster 1, H2ac	19	17	16	1.95E-05	23
HIST1H2AM	histone cluster 1, H2am	18	17	13	3.80E-03	26
HIST1H2BC	histone cluster 1, H2bc	25	18	20	7.68E-06	22
HIST1H2BK	histone cluster 1, H2bk	27	24	27	1.71E-04	24
HIST1H3B	histone cluster 1, H3b	10	10	6	8.60E-02	29
HIST2H2BE	histone cluster 2, H2be	17	14	13	8.99E-02	29
HLA-A	major histocompatibility complex, class I, A	27	25	24	1.15E-12	16
HLA-B	major histocompatibility complex, class I, B	36	35	26	2.02E-13	14
HLA-C	major histocompatibility complex, class I, C	16	13	12	8.88E-11	18
HLA-DMA	major histocompatibility complex, class II, DM alpha	8	6	8	1.94E-02	27
HVCN1	hydrogen voltage-gated channel 1	11	10	10	8.95E-07	21
IGLL5	immunoglobulin lambda-like polypeptide 5	51	30	39	3.70E-08	20
IKZF3	IKAROS family zinc finger 3 (Aiolos)	11	10	8	4.97E-06	22
IL6	interleukin 6 (interferon, beta 2)	6	6	6	9.82E-02	30
IRF8	interferon regulatory factor 8	25	24	20	6.47E-04	24
KLHL6	kelch-like 6 (Drosophila)	37	28	31	2.02E-13	15
KMT2D	myeloid/lymphoid or mixed-lineage leukemia 2	96	75	95	2.07E-11	18
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog	8	8	3	4.97E-06	22
LTB	lymphotoxin beta (TNF superfamily, member 3)	29	20	23	5.48E-03	26
LYN	v-yes-1 Yamaguchi sarcoma viral related oncogene homolog	15	12	14	6.11E-07	21
MEF2B	myocyte enhancer factor 2B	20	20	19	2.99E-11	18
MYD88	myeloid differentiation primary response gene (88)	57	55	10	2.02E-13	14
NANOG	Nanog homeobox	6	5	4	1.22E-02	27
NAV1	neuron navigator 1	7	7	5	3.46E-03	26

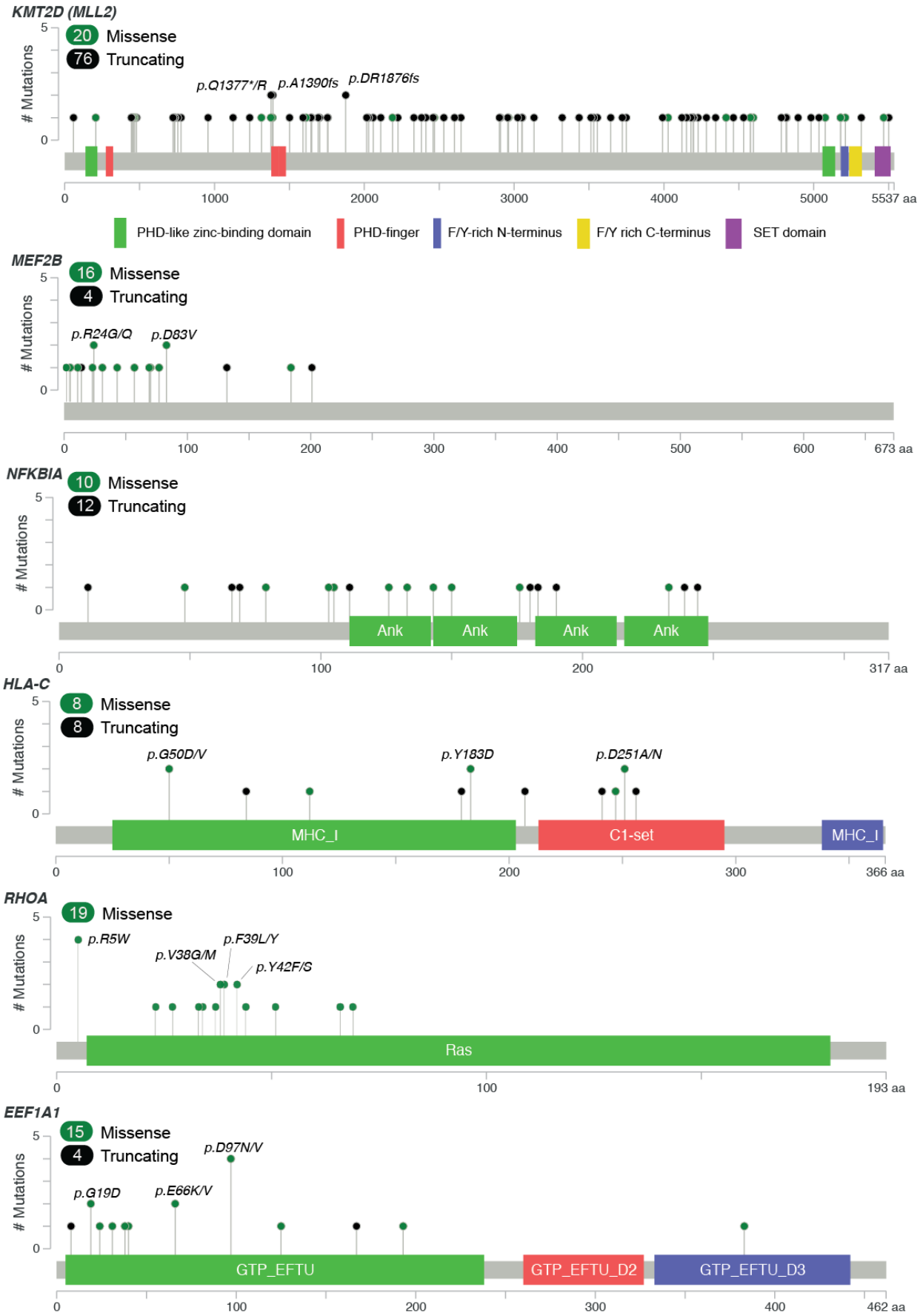
NFKBIA	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha	22	15	22	5.22E-11	18
NFKBIE	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, epsilon	12	10	9	1.43E-04	24
NLRP8	NLR family, pyrin domain containing 8	10	10	9	6.37E-02	28
NOTCH2	Notch homolog 2 (Drosophila)	21	20	15	1.00E-06	21
PDE4DIP	phosphodiesterase 4D interacting protein (myomegalin)	27	24	24	3.49E-02	28
PIM1	pim-1 oncogene	142	67	84	1.35E-12	16
POU2AF1	POU class 2 associating factor 1	12	12	8	4.01E-10	19
POU2F2	POU class 2 homeobox 2	17	17	9	8.54E-09	19
PRDM1	PR domain containing 1, with ZNF domain	23	22	19	2.02E-13	15
PRKCB	protein kinase C, beta	12	11	10	3.14E-02	27
PRPS1	phosphoribosyl pyrophosphate synthetase 1	4	4	3	3.26E-02	27
PTEN	phosphatase and tensin homolog (mutated in multiple advanced cancers 1)	11	10	10	3.88E-05	23
PTPN6	protein tyrosine phosphatase, non-receptor type 6	21	13	20	2.17E-05	23
RAD9A	RAD9 homolog A (S. pombe)	5	5	3	2.08E-03	25
RHOA	ras homolog gene family, member A	19	16	15	1.03E-10	18
SF3B1	splicing factor 3b, subunit 1, 155kDa	9	9	7	5.05E-02	28
SGK1	serum/glucocorticoid regulated kinase 1	118	43	73	3.13E-08	20
SIN3A	SIN3 homolog A, transcription regulator (yeast)	12	11	12	2.46E-02	27
SMEK1	SMEK homolog 1, suppressor of mek1 (Dictyostelium)	8	8	8	6.60E-03	26
SPEN	spen homolog, transcriptional regulator (Drosophila)	29	27	28	6.87E-04	25
STAT3	signal transducer and activator of transcription 3 (acute-phase response factor)	22	19	18	4.72E-08	20
STAT6	signal transducer and activator of transcription 6, interleukin-4 induced	16	14	11	5.24E-07	21
TBL1XR1	transducin (beta)-like 1 X-linked receptor 1	23	22	18	2.02E-13	15
TLR2	toll-like receptor 2	9	9	7	3.96E-02	28
TMEM30A	transmembrane protein 30A	19	17	16	2.22E-12	17
TMSB4X	thymosin beta 4, X-linked	44	38	27	9.32E-12	17
TNFAIP3	tumor necrosis factor, alpha-induced protein 3	29	26	28	1.45E-11	17
TNFRSF14	tumor necrosis factor receptor superfamily, member 14 (herpesvirus entry mediator)	43	42	39	2.22E-12	16
TOX	thymocyte selection-associated high mobility group box	13	12	9	6.83E-06	22
TP53	tumor protein p53	70	65	53	2.02E-13	14
UBE2A	ubiquitin-conjugating enzyme E2A (RAD6 homolog)	13	12	12	2.88E-10	19
XPO1	exportin 1 (CRM1 homolog, yeast)	7	7	3	5.39E-04	24
YY1	YY1 transcription factor	8	8	8	2.53E-03	26
ZC3H12A	zinc finger CCCH-type containing 12A	10	10	10	3.30E-04	24
ZEB2	zinc finger E-box binding homeobox 2	14	13	14	8.00E-02	29
ZFP36L1	zinc finger protein 36, C3H type-like 1	30	25	20	1.73E-03	25
ZNF423	zinc finger protein 423	5	5	5	4.80E-02	28

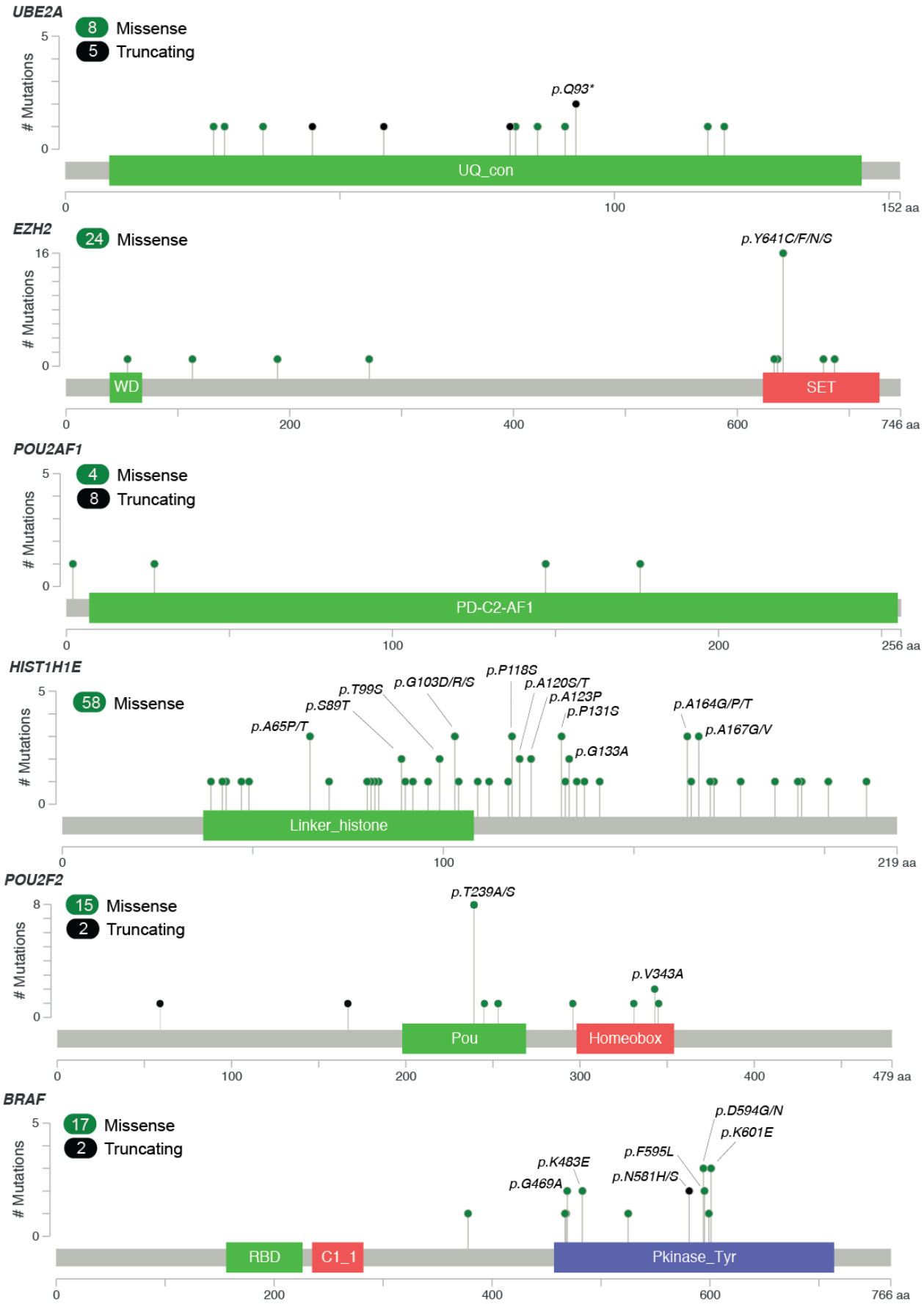

TP53
54 Missense
16 Truncating
Mutations
p.C176F/G/R/S
p.R175H
p.Y205C
p.R213*
p.N239D/S
p.M237I
p.R248Q/W
p.G245C/S/V
p.R282G/W
Trans-activation
DNA binding domain
Tetramer-isation
0
100
200
300
393 aa
MYD88
55 Missense
2 Truncating
p.L265P
p.S243N
p.S219C
p.M232T
p.V217F
Death
TIR
296 aa
CREBBP
39 Missense
15 Truncating
1 In-Frame
p.R1446C/H
p.D1435G/N/V
p.Y1503C/H
p.R1360*
p.Q1491K/P
zf-TAZ
KIX
Bromo
DUF
HAT_KAT11
ZZ
zf-TAZ
Creb binding
400
800
1200
1600
2000
2442 aa
CD79B
39 Missense
10 Truncating
1 In-Frame
p.Y196C/196D/196F/196H/196N/196S/196del/E196del
p.E197A/G
p.L199P/Q
p.-206fs/T206fs
p.Y207*/H
p.K183E/M
V-set
229 aa
HLA-B
19 Missense
17 Truncating
p.G50D
p.A141T/V
p.Q78*
p.Y147D/N
p.Q286*
p.H287N/Y
MHC_I
C1-set
MHC_I
362 aa
CD70
16 Missense
22 Truncating
p.G66*/R
p.Q47*
p.R83C/H
p.C151S/fs
p.W190*
TNF
193 aa

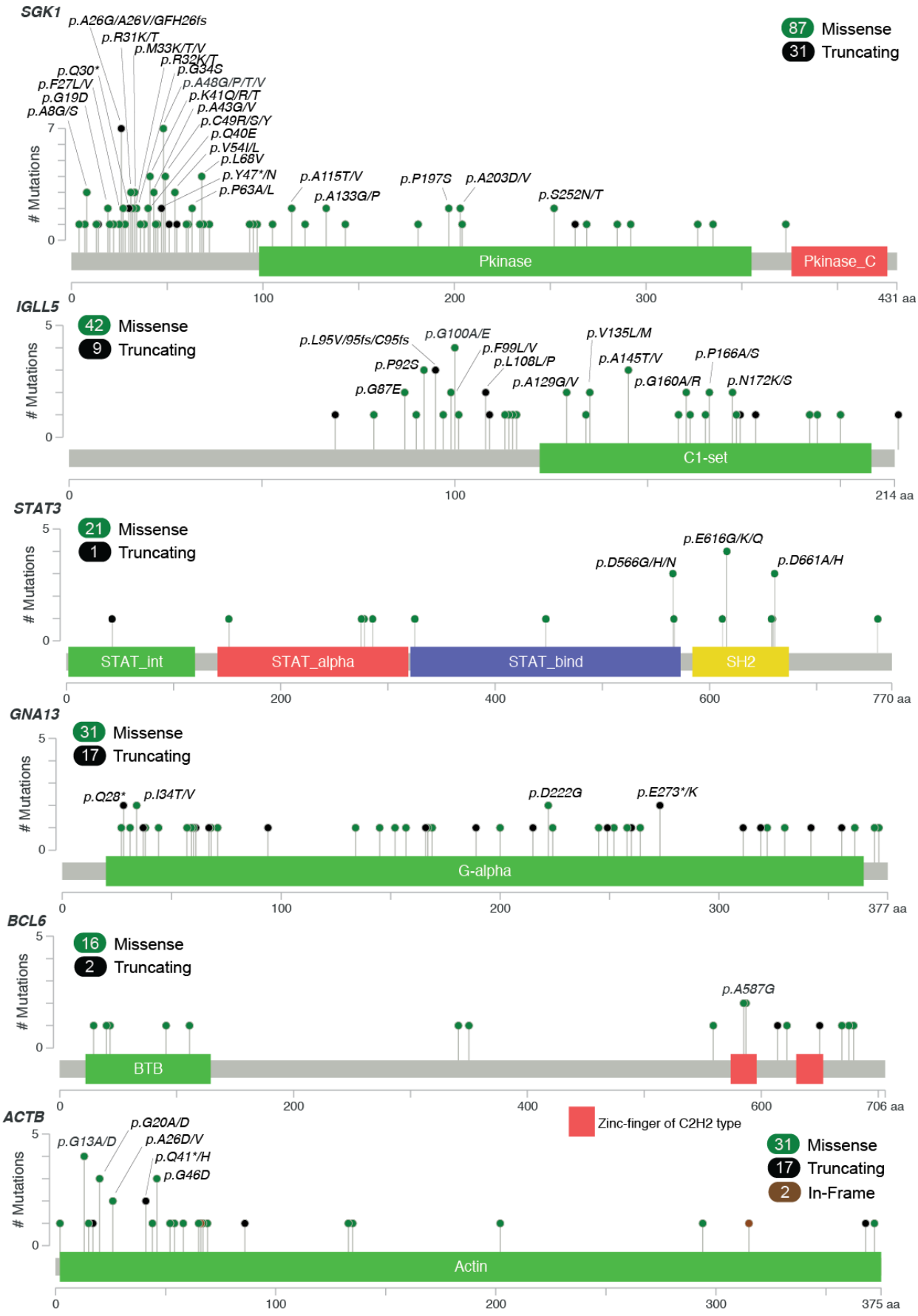

FAS
9 Missense
18 Truncating
Mutations
p.V220fs
p.R250*/G
TNFR_c6
TNFR_c6
Death
0
100
200
300
335 aa
KLHL6
37 Missense
p.L65P/Q
p.L90P/R
p.N69D/S
p.E547K
BTB
BACK
Kelch_1
Kelch_1
Kelch_1
621 aa
PRDM1
8 Missense
15 Truncating
p.E97E
SET
zf-C2H2
zf-C2H2
825 aa
TBL1XR1
18 Missense
3 Truncating
2 In-Frame
p.Y395C/S
p.Y446C
LisH
WD40
WD40
WD40
WD40
WD40
WD40
514 aa
BTG1
61 Missense
7 Truncating
p.E46D/G/K/Q/V
p.P3L/R
p.T6I
p.L37M/37V/Q37fs
p.M11I/L
p.S23A/C
p.L31F
p.L47Q/V
p.I70L/T
p.V111E/M
BTG
171 aa
BCL2
111 Missense
p.A43G/V
p.A2P/S
p.G33A/R
p.P46A/S
p.R6G/T
p.G47D/V
p.T7R/S
p.P59A/L/S
p.M16L/V
p.D34H
p.Q73E/L
p.T74I/P
p.H20Q
p.L86F/V
p.P90S
p.Q118H/R
p.S117R
p.H120R/Y
p.A113D/G
p.F124I/L/V
p.Y108C/H
p.R129H/S
p.A131D/G/V
p.N172S
p.N163K/S
p.Y180D/F
p.H186Q/Y
BH4
Bcl-2
239 aa

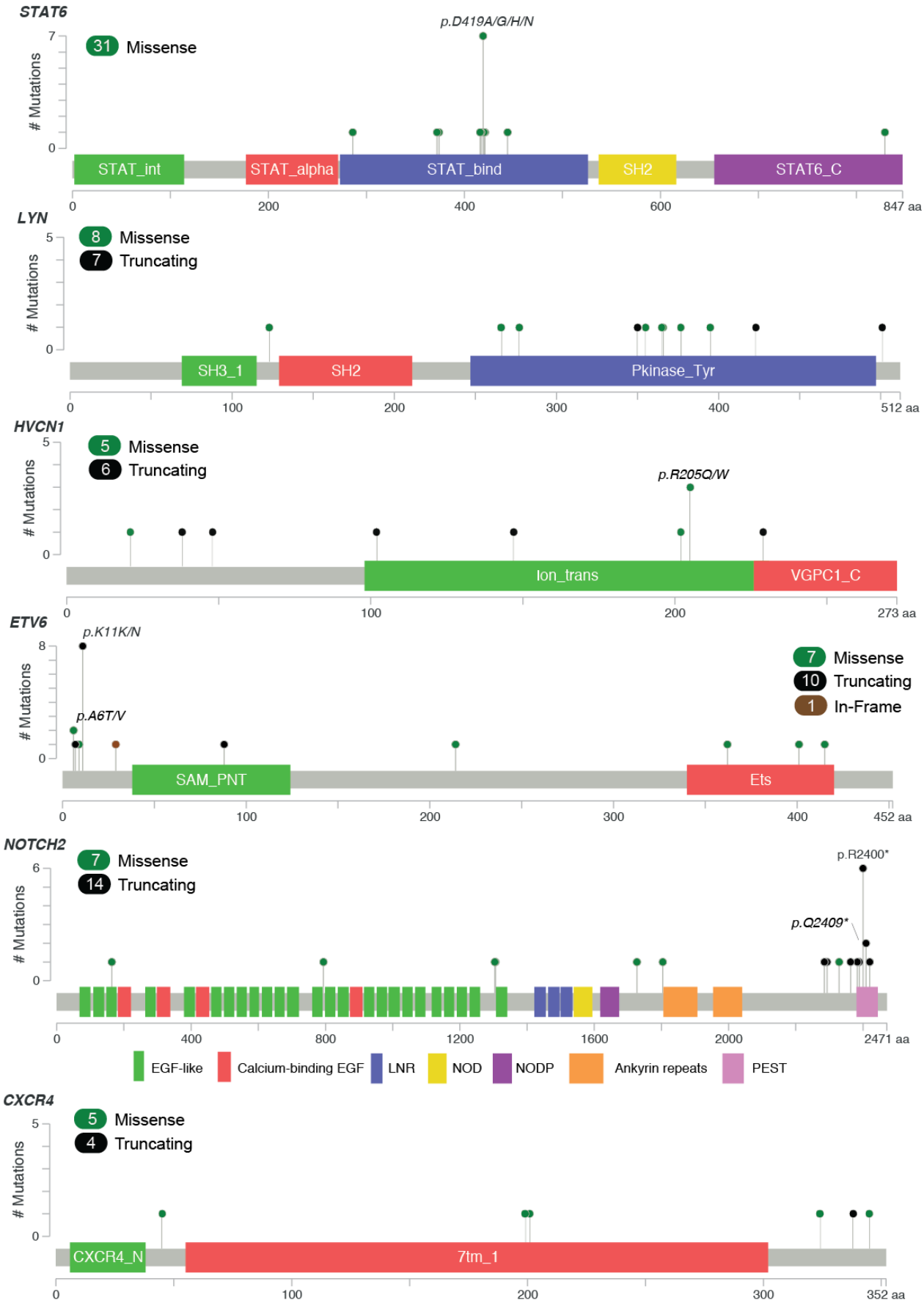

CARD11
42 Missense
1 Truncating
p.G123D/S
p.F115I/L
p.R114G/T
p.S250P
p.L251P
p.D357V
p.E626K
Mutations
CARD
0
200
400
600
800
1000
1154 aa
HLA-A
10 Missense
16 Truncating
1 In-Frame
p.G50D/V
p.Q78*
MHC_I
C1-set
MHC_I
365 aa
PIM1
114 Missense
28 Truncating
p.L2F/V
p.A10D/V
p.E70*/D/K/Q
p.C17G/Y
p.K24N
p.L25V/APG25fs
p.G28D
p.E30K
p.P33S/T
p.Q37H
p.Q39*/E/H
p.S59F/Y
p.E79D/Q
p.P81L/S/T
p.S75C/F
p.D72G/N
p.G83D/V
p.S97N/R/T
p.G99D
p.S101F
p.L106F
p.L120M/V
p.P125A/S/T
p.Q127*/E
p.V126L/M
p.L164F/V
p.L184F/V
p.E135Q/fs
p.S146N/R
p.K183M/N
p.L193F
Pkinase
404 aa
CCND3
6 Missense
8 Truncating
p.R271fs
p.P284L/S
p.L292R/fs
Cyclin_N
Cyclin_C
292 aa
B2M
11 Missense
20 Truncating
p.L7*/S
p.L12P/R
p.L13R/fs
p.N41K
p.-80fs/W80*
p.Y83C/fs
p.Y98C/S
C1-set
119 aa
TNFRSF14
19 Missense
23 Truncating
1 In-Frame
p.W12*
p.Y26D
p.C42F/S
p.Y61*/D
p.W204*
TNFR_c6
283 aa

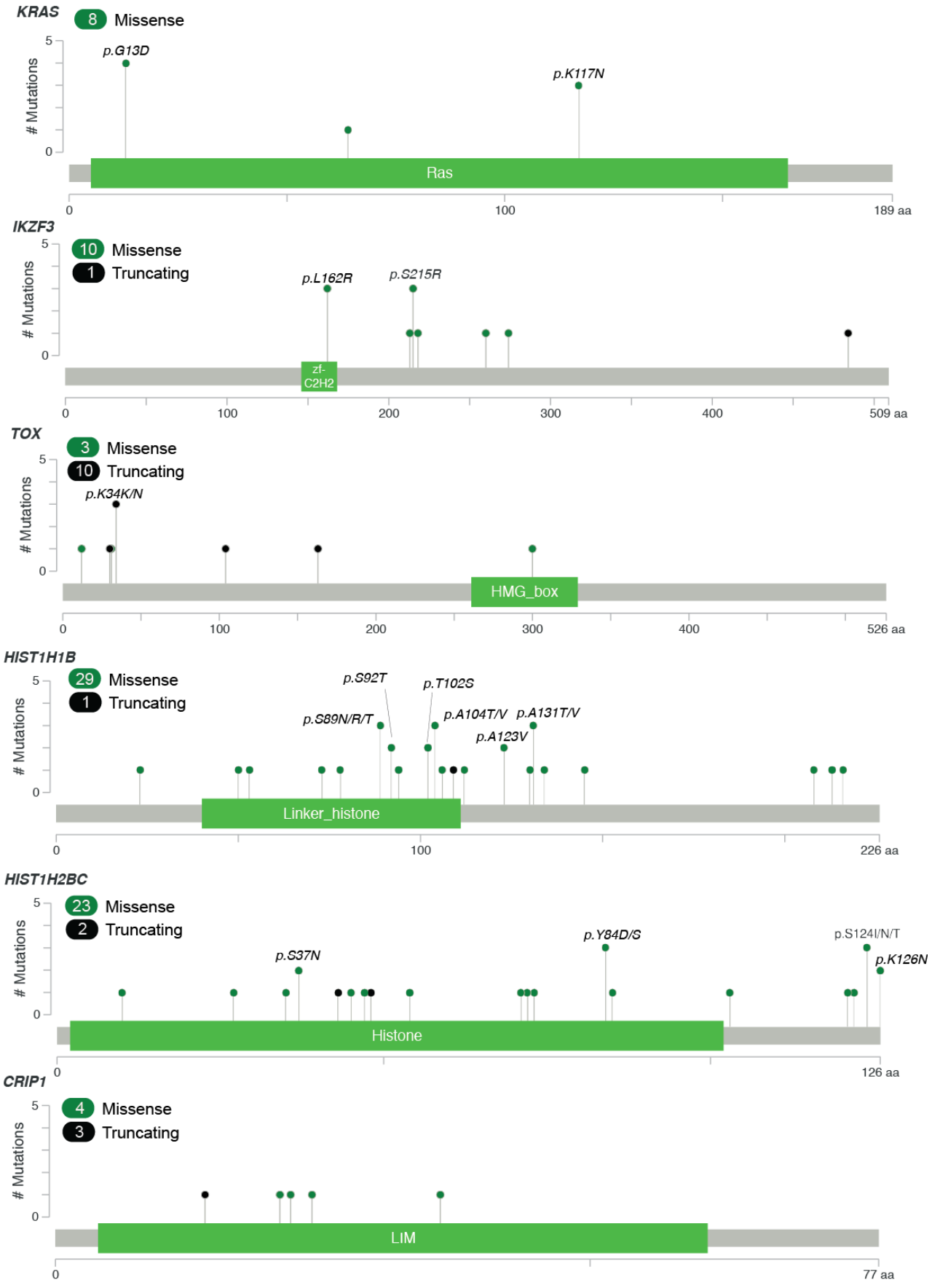

HIST1H1C
39 Missense
2 Truncating
Mutations
p.E53D/Q
p.V57L
p.A61G/T
p.A65V
p.S78N/R
p.S89T
p.A101D/G/V
p.S150N/T
p.A171T
p.S173R/T
Linker_histone
0
100
213 aa
TMEM30A
3 Missense
16 Truncating
p.R161*
p.R226*
CDC50
0
100
200
300
361 aa
TMSB4X
23 Missense
20 Truncating
1 In-Frame
p.P5S/fs
p.I10S/T
p.E22K/fs
p.Q24*
p.N27S/Y
p.P28L/Q/S
p.E33D/K
p.Q40*
Thymosin
0
44 aa
EBF1
21 Missense
16 Truncating
p.W30*
p.N66K
p.C170G
p.H235Q/Y
COE1_DBD
TIG
COE..
0
100
200
300
400
500
591 aa
CD58
10 Missense
20 Truncating
p.W30*
p.S212*
0
100
200
250 aa
TNFAIP3 (A20)
2 Missense
22 Truncating
p.D134fs
OTU
zf-A20
zf-A20
zf-A20
zf-A20
0
200
400
600
790 aa

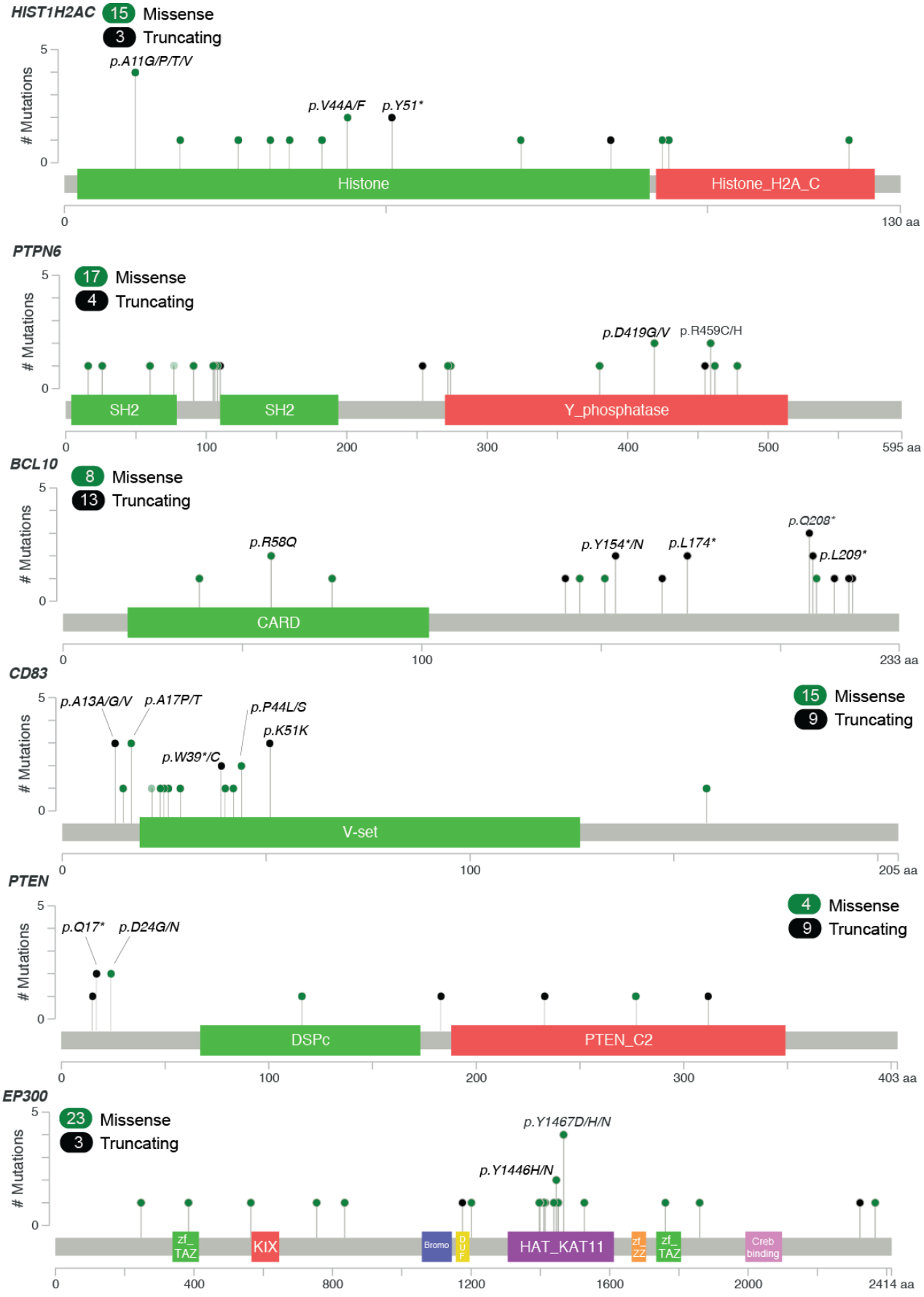

KMT2D (MLL2)
20 Missense
76 Truncating
p.Q1377*/R
p.A1390fs
p.DR1876fs
Mutations
5
0
0
1000
2000
3000
4000
5000
5537 aa
PHD-like zinc-binding domain
PHD-finger
F/Y-rich N-terminus
F/Y rich C-terminus
SET domain
MEF2B
16 Missense
4 Truncating
p.R24G/Q
p.D83V
Mutations
5
0
0
100
200
300
400
500
600
673 aa
NFKBIA
10 Missense
12 Truncating
Mutations
5
0
Ank
Ank
Ank
Ank
0
100
200
317 aa
HLA-C
8 Missense
8 Truncating
p.G50D/V
p.Y183D
p.D251A/N
Mutations
5
0
MHC_I
C1-set
MHC_I
0
100
200
300
366 aa
RHOA
19 Missense
p.R5W
p.V38G/M
p.F39L/Y
p.Y42F/S
Mutations
5
0
Ras
0
100
193 aa
EEF1A1
15 Missense
4 Truncating
p.D97N/V
p.G19D
p.E66K/V
Mutations
5
0
GTP_EFTU
GTP_EFTU_D2
GTP_EFTU_D3
0
100
200
300
400
462 aa

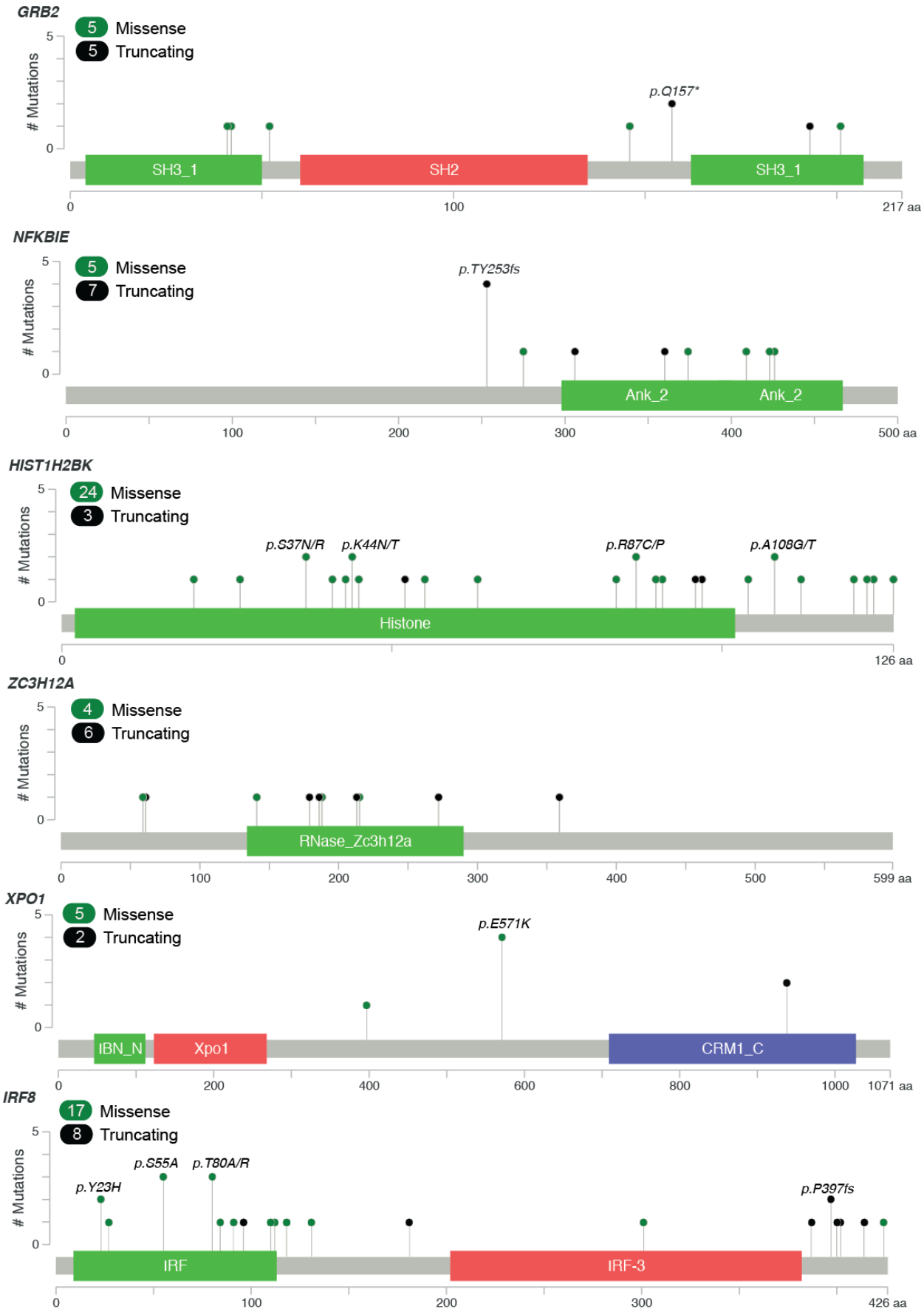

UBE2A
8 Missense
5 Truncating
Mutations
p.Q93*
UQ_con
0
100
152 aa
EZH2
24 Missense
16
Mutations
p.Y641C/F/N/S
WD
SET
0
200
400
600
746 aa
POU2AF1
4 Missense
8 Truncating
Mutations
PD-C2-AF1
0
100
200
256 aa
HIST1H1E
58 Missense
Mutations
p.T99S
p.G103D/R/S
p.P118S
p.A120S/T
p.A123P
p.S89T
p.P131S
p.A164G/P/T
p.A65P/T
p.A167G/V
p.G133A
Linker_histone
0
100
219 aa
POU2F2
15 Missense
2 Truncating
Mutations
p.T239A/S
p.V343A
Pou
Homeobox
0
100
200
300
400
479 aa
BRAF
17 Missense
2 Truncating
Mutations
p.D594G/N
p.K601E
p.F595L
p.K483E
p.N581H/S
p.G469A
RBD
C1_1
Pkinase_Tyr
0
200
400
600
766 aa

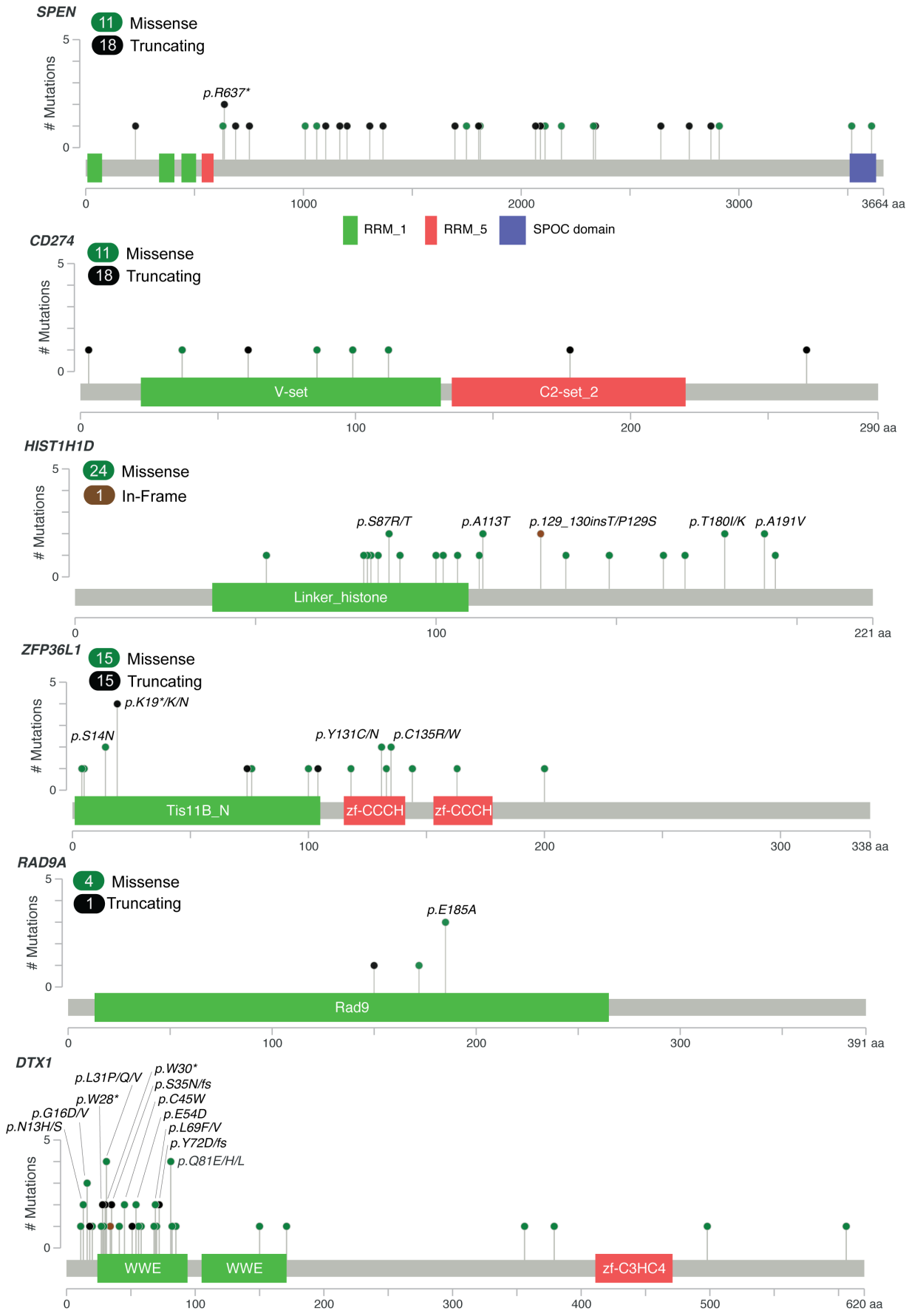

SGK1
87 Missense
31 Truncating
p.A26G/A26V/GFH26fs
p.R31K/T
p.M33K/T/V
p.R32K/T
p.G34S
p.A48G/P/T/V
p.K41Q/R/T
p.A43G/V
p.C49R/S/Y
p.Q40E
p.V54I/L
p.L68V
p.Y47*/N
p.P63A/L
p.Q30*
p.F27L/V
p.G19D
p.A8G/S
p.A115T/V
p.A133G/P
p.P197S
p.A203D/V
p.S252N/T
Mutations
Pkinase
Pkinase_C
431 aa
IGLL5
42 Missense
9 Truncating
p.L95V/95fs/C95fs
p.G100A/E
p.F99L/V
p.L108L/P
p.P92S
p.G87E
p.A129G/V
p.V135L/M
p.A145T/V
p.G160A/R
p.P166A/S
p.N172K/S
C1-set
214 aa
STAT3
21 Missense
1 Truncating
p.E616G/K/Q
p.D566G/H/N
p.D661A/H
STAT_int
STAT_alpha
STAT_bind
SH2
770 aa
GNA13
31 Missense
17 Truncating
p.Q28*
p.I34T/V
p.D222G
p.E273*/K
G-alpha
377 aa
BCL6
16 Missense
2 Truncating
p.A587G
BTB
706 aa
Zinc-finger of C2H2 type
ACTB
31 Missense
17 Truncating
2 In-Frame
p.G20A/D
p.G13A/D
p.A26D/V
p.Q41*/H
p.G46D
Actin
375 aa

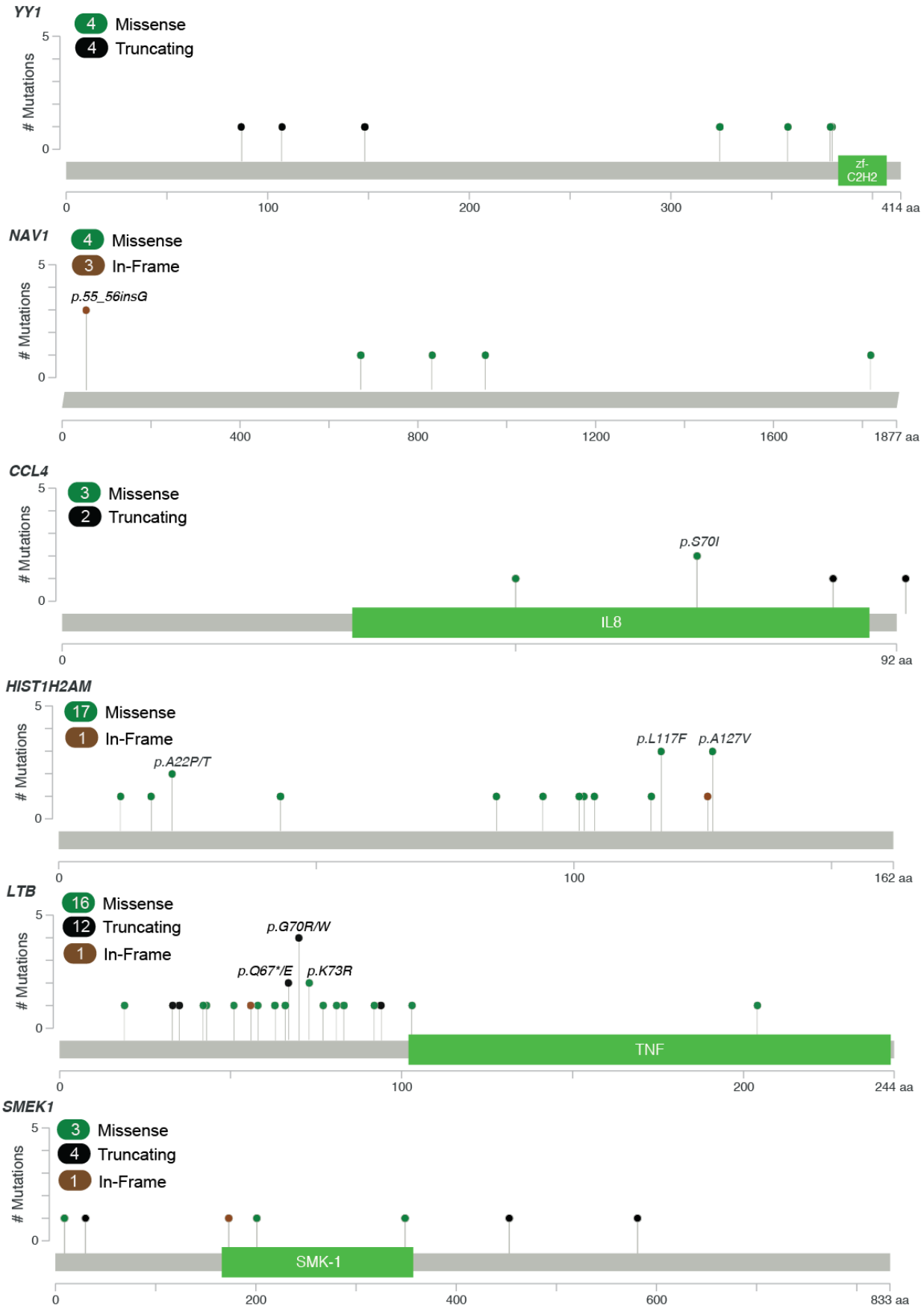

STAT6
p.D419A/G/H/N
31 Missense
Mutations
STAT_int
STAT_alpha
STAT_bind
SH2
STAT6_C
847 aa
LYN
8 Missense
7 Truncating
SH3_1
SH2
Pkinase_Tyr
512 aa
HVCN1
5 Missense
6 Truncating
p.R205Q/W
Ion_trans
VGPC1_C
273 aa
ETV6
p.K11K/N
p.A6T/V
7 Missense
10 Truncating
1 In-Frame
SAM_PNT
Ets
452 aa
NOTCH2
7 Missense
14 Truncating
p.R2400*
p.Q2409*
2471 aa
EGF-like
Calcium-binding EGF
LNR
NOD
NODP
Ankyrin repeats
PEST
CXCR4
5 Missense
4 Truncating
CXCR4_N
7tm_1
352 aa

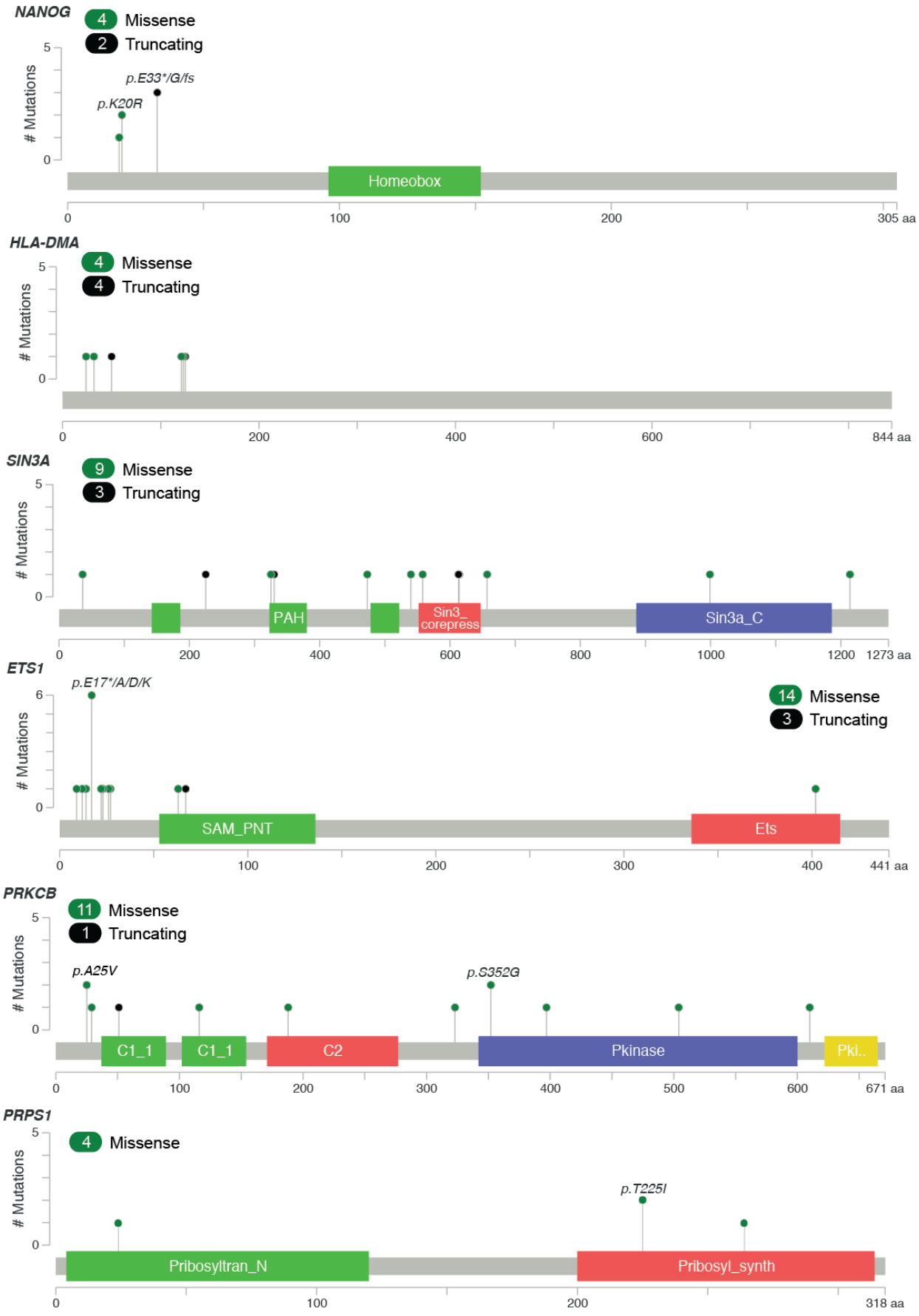

KRAS
8 Missense
p.G13D
p.K117N
Mutations
Ras
0
100
189 aa
IKZF3
10 Missense
1 Truncating
p.L162R
p.S215R
zf-C2H2
200
300
400
509 aa
TOX
3 Missense
10 Truncating
p.K34K/N
HMG_box
526 aa
HIST1H1B
29 Missense
1 Truncating
p.S92T
p.T102S
p.S89N/R/T
p.A104T/V
p.A131T/V
p.A123V
Linker_histone
226 aa
HIST1H2BC
23 Missense
2 Truncating
p.S37N
p.Y84D/S
p.S124I/N/T
p.K126N
Histone
126 aa
CRIP1
4 Missense
3 Truncating
LIM
77 aa

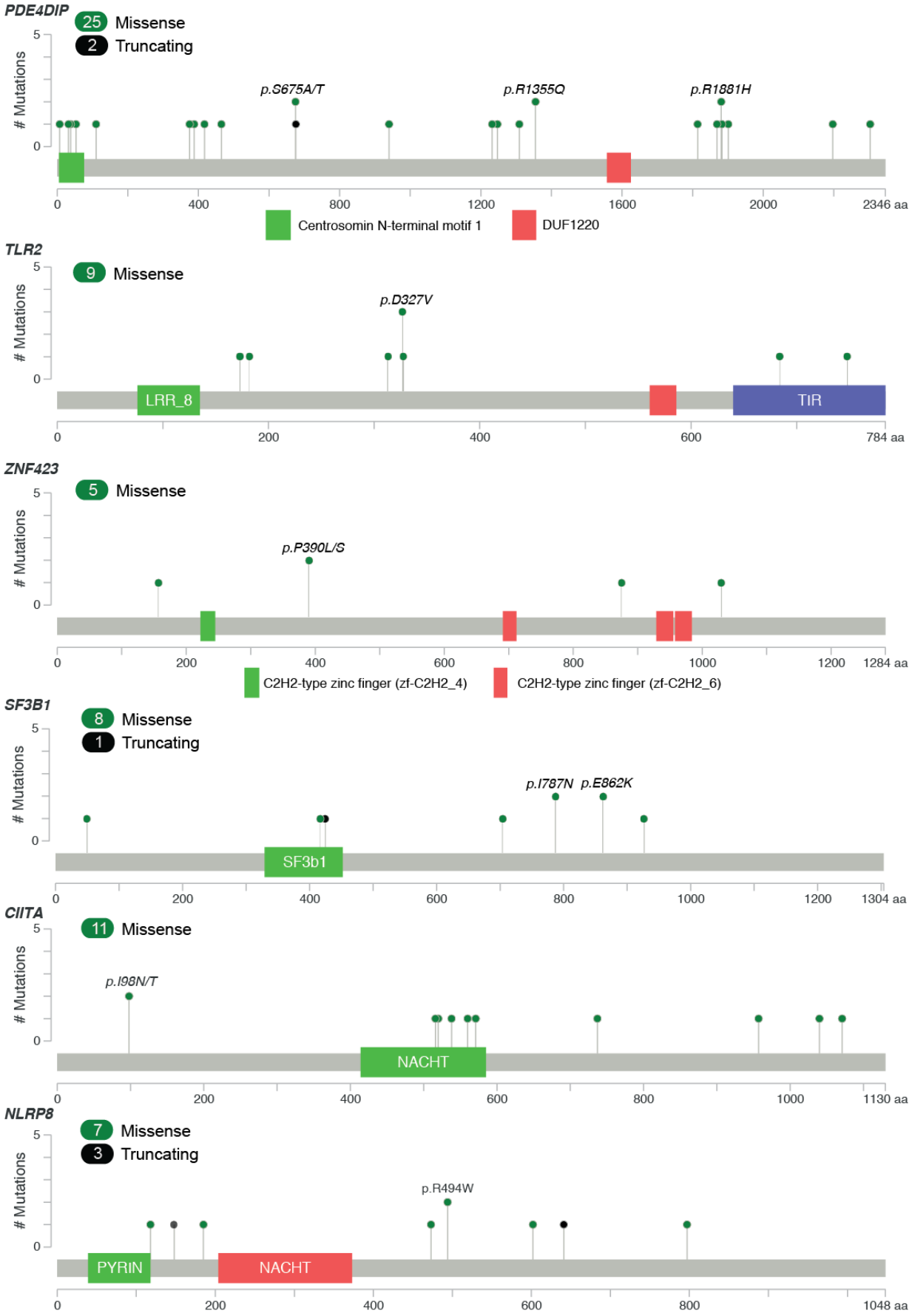

HIST1H2AC
15 Missense
3 Truncating
p.A11G/P/T/V
p.V44A/F
p.Y51*
Mutations
Histone
Histone_H2A_C
130 aa
PTPN6
17 Missense
4 Truncating
p.D419G/V
p.R459C/H
SH2
SH2
Y_phosphatase
595 aa
BCL10
8 Missense
13 Truncating
p.R58Q
p.Y154*/N
p.L174*
p.Q208*
p.L209*
CARD
233 aa
CD83
15 Missense
9 Truncating
p.A13A/G/V
p.A17P/T
p.P44L/S
p.K51K
p.W39*/C
V-set
205 aa
PTEN
4 Missense
9 Truncating
p.Q17*
p.D24G/N
DSPc
PTEN_C2
403 aa
EP300
23 Missense
3 Truncating
p.Y1467D/H/N
p.Y1446H/N
zf-TAZ
KIX
Bromo
DUF
HAT_KAT11
zf-ZZ
zf-TAZ
Creb binding
2414 aa

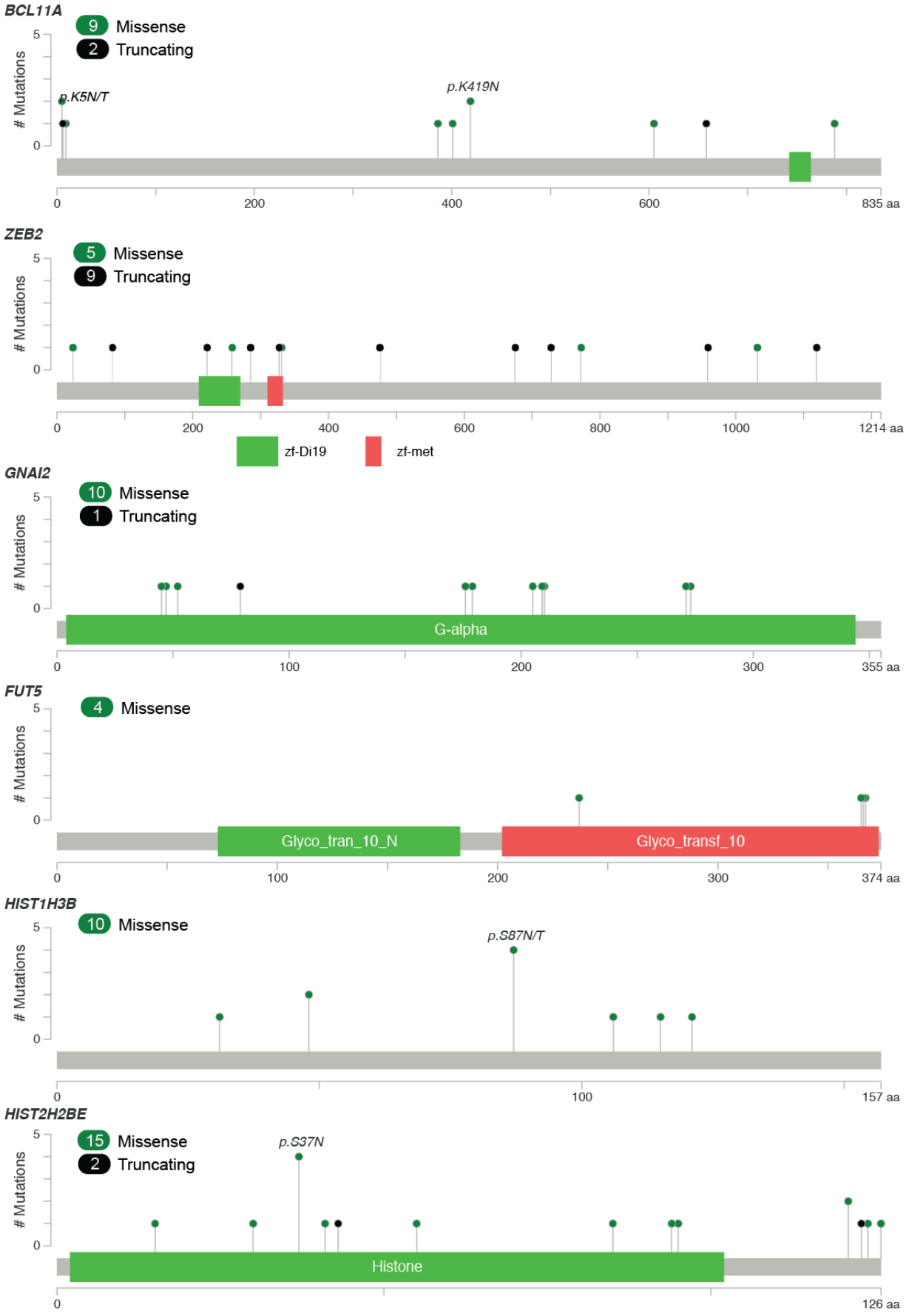

GRB2
5 Missense
5 Truncating
Mutations
p.Q157*
SH3_1
SH2
SH3_1
217 aa
NFKBIE
5 Missense
7 Truncating
p.TY253fs
Ank_2
Ank_2
500 aa
HIST1H2BK
24 Missense
3 Truncating
p.S37N/R
p.K44N/T
p.R87C/P
p.A108G/T
Histone
126 aa
ZC3H12A
4 Missense
6 Truncating
RNase_Zc3h12a
599 aa
XPO1
5 Missense
2 Truncating
p.E571K
IBN_N
Xpo1
CRM1_C
1071 aa
IRF8
17 Missense
8 Truncating
p.S55A
p.T80A/R
p.Y23H
p.P397fs
IRF
IRF-3
426 aa

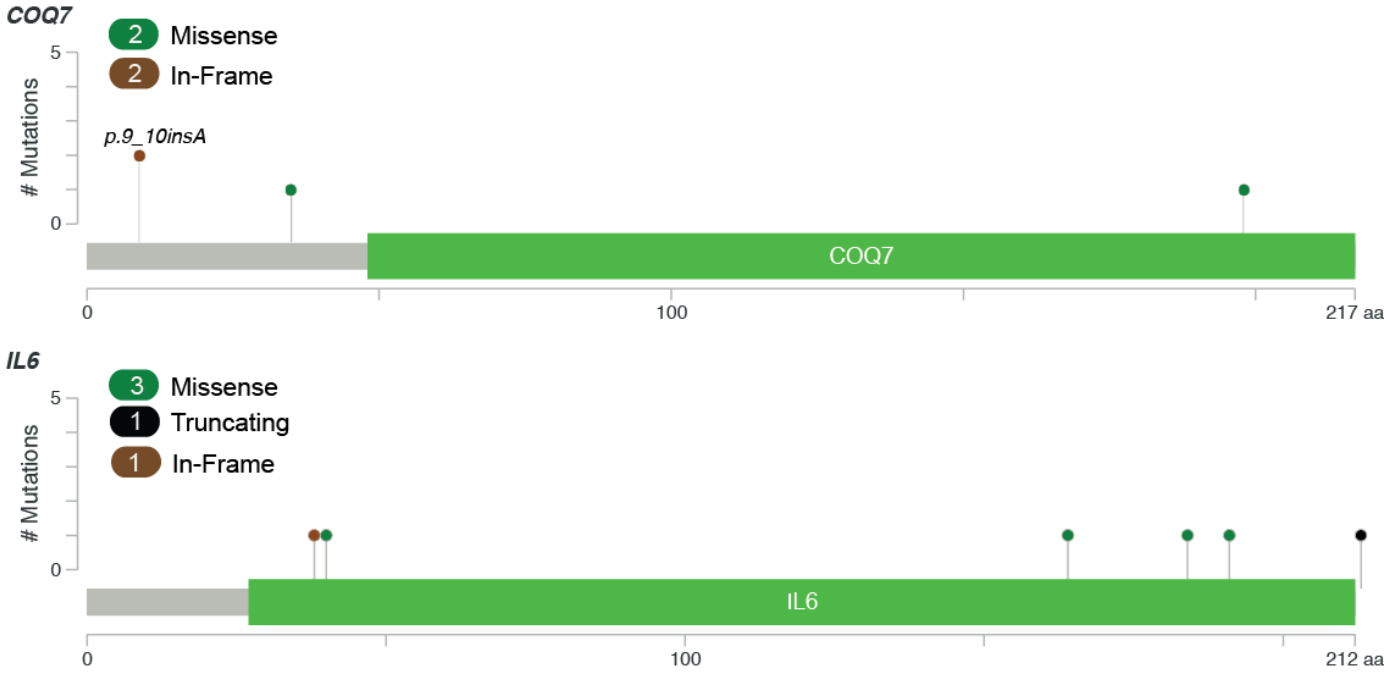

SPEN
11 Missense
18 Truncating
Mutations
p.R637*
RRM_1
RRM_5
SPOC domain
3664 aa
CD274
11 Missense
18 Truncating
V-set
C2-set_2
290 aa
HIST1H1D
24 Missense
1 In-Frame
p.S87R/T
p.A113T
p.129_130insT/P129S
p.T180I/K
p.A191V
Linker_histone
221 aa
ZFP36L1
15 Missense
15 Truncating
p.K19*/K/N
p.S14N
p.Y131C/N
p.C135R/W
Tis11B_N
zf-CCCH
zf-CCCH
338 aa
RAD9A
4 Missense
1 Truncating
p.E185A
Rad9
391 aa
DTX1
p.L31P/Q/V
p.W30*
p.S35N/fs
p.W28*
p.C45W
p.G16D/V
p.E54D
p.N13H/S
p.L69F/V
p.Y72D/fs
p.Q81E/H/L
WWE
WWE
zf-C3HC4
620 aa


YY1
4 Missense
4 Truncating
Mutations
zf-C2H2
0 100 200 300 414 aa
NAV1
4 Missense
3 In-Frame
p.55_56insG
0 400 800 1200 1600 1877 aa
CCL4
3 Missense
2 Truncating
p.S70I
IL8
0 92 aa
HIST1H2AM
17 Missense
1 In-Frame
p.A22P/T
p.L117F
p.A127V
0 100 162 aa
LTB
16 Missense
12 Truncating
1 In-Frame
p.G70R/W
p.Q67*/E
p.K73R
TNF
0 100 200 244 aa
SMEK1
3 Missense
4 Truncating
1 In-Frame
SMK-1
0 200 400 600 833 aa

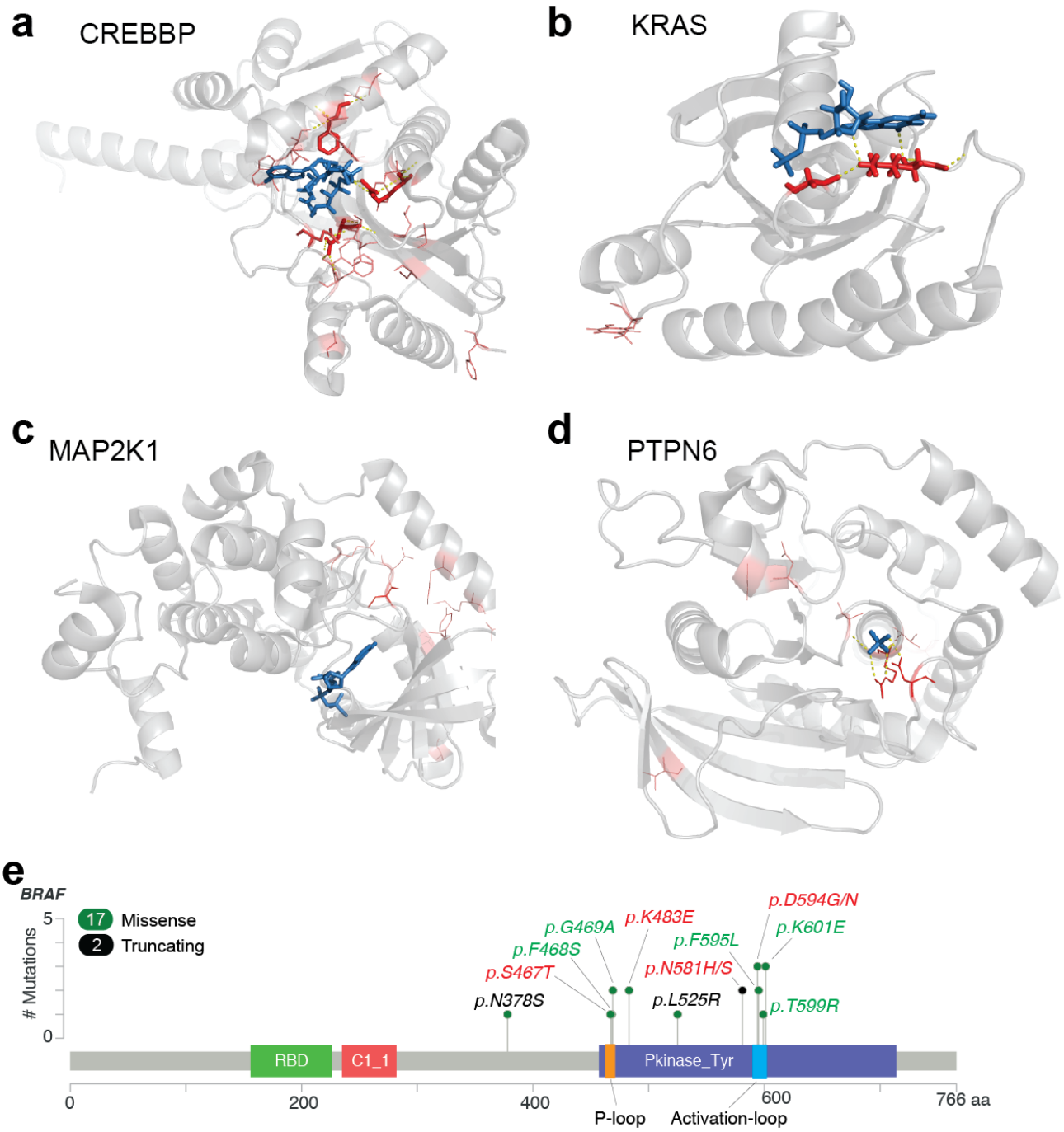

NANOG
4 Missense
2 Truncating
Mutations
p.E33*/G/fs
p.K20R
Homeobox
0
100
200
305 aa
HLA-DMA
4 Missense
4 Truncating
844 aa
SIN3A
9 Missense
3 Truncating
PAH
Sin3_ corepress
Sin3a_C
1273 aa
ETS1
p.E17*/A/D/K
14 Missense
3 Truncating
SAM_PNT
Ets
441 aa
PRKCB
11 Missense
1 Truncating
p.A25V
p.S352G
C1_1
C1_1
C2
Pkinase
Pki..
671 aa
PRPS1
4 Missense
p.T225I
Pribosyltran_N
Pribosyl_synth
318 aa


PDE4DIP
25 Missense
2 Truncating
Mutations
p.S675A/T
p.R1355Q
p.R1881H
Centrosomin N-terminal motif 1
DUF1220
2346 aa
TLR2
9 Missense
p.D327V
LRR_8
TIR
784 aa
ZNF423
5 Missense
p.P390L/S
C2H2-type zinc finger (zf-C2H2_4)
C2H2-type zinc finger (zf-C2H2_6)
1284 aa
SF3B1
8 Missense
1 Truncating
p.I787N p.E862K
SF3b1
1304 aa
CIITA
11 Missense
p.I98N/T
NACHT
1130 aa
NLRP8
7 Missense
3 Truncating
p.R494W
PYRIN
NACHT
1048 aa

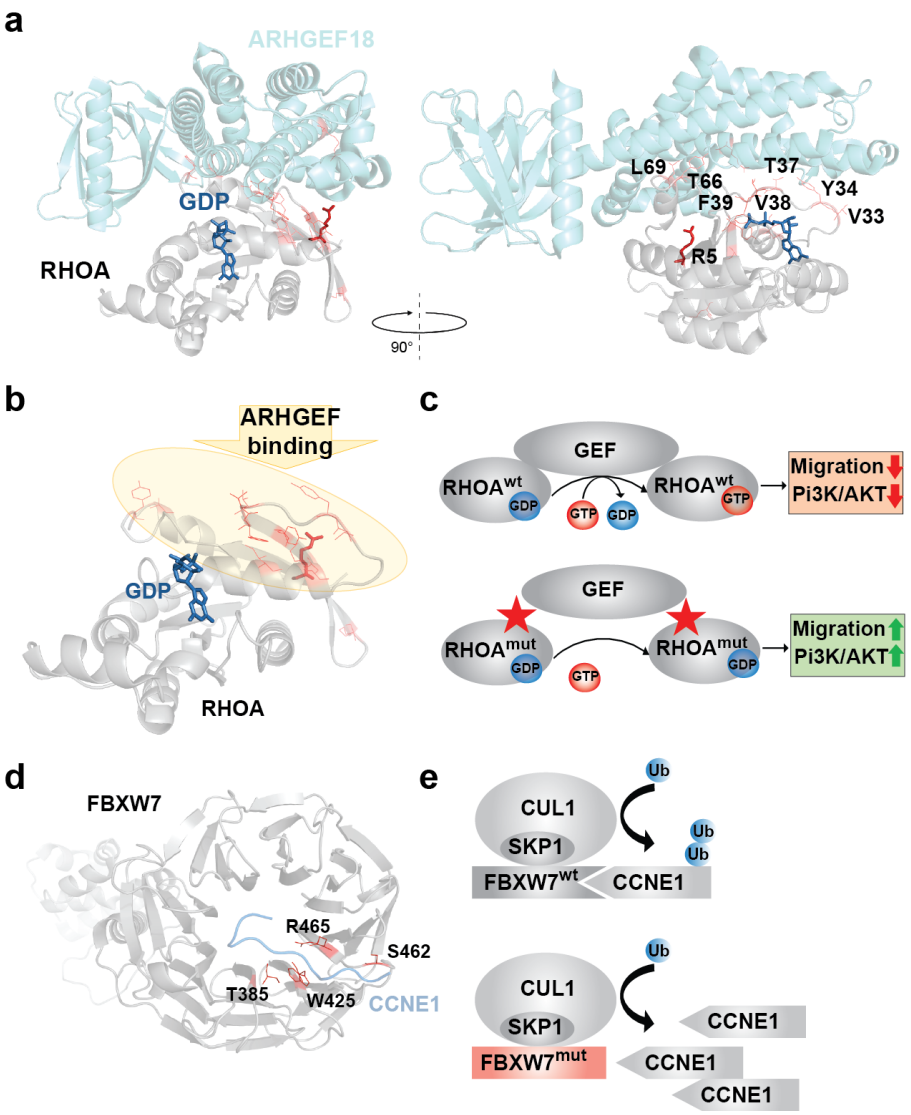

BCL11A
9 Missense
2 Truncating
Mutations
p.K5N/T
p.K419N
0
200
400
600
835 aa
ZEB2
5 Missense
9 Truncating
1000
1214 aa
zf-Di19
zf-met
GNAI2
10 Missense
1 Truncating
G-alpha
100
300
355 aa
FUT5
4 Missense
Glyco_tran_10_N
Glyco_transf_10
374 aa
HIST1H3B
10 Missense
p.S87N/T
157 aa
HIST2H2BE
15 Missense
2 Truncating
p.S37N
Histone
126 aa



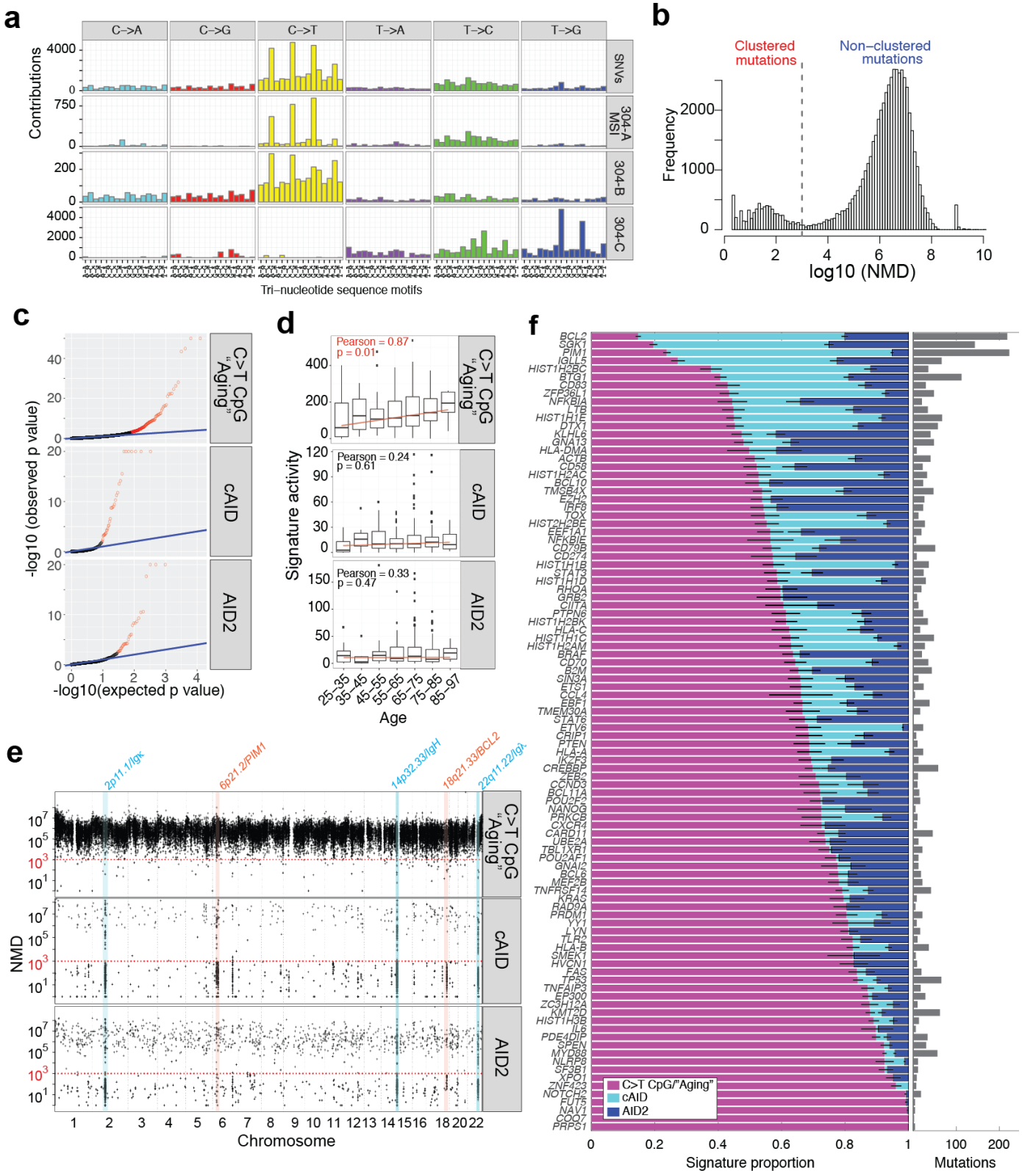
Supplementary Figure 4: Mutation diagrams (lollipop figures) for all significantly mutated genes. For each significantly mutated gene, all non-synonymous mutations are visualized within the functional domains of the respective protein using *MutationMapper* v1.0.1[8,9]. Genes are ordered by significance (*MutSig2CV* q-value). An alphabetic index of all mutation diagrams is included on page 12/13.

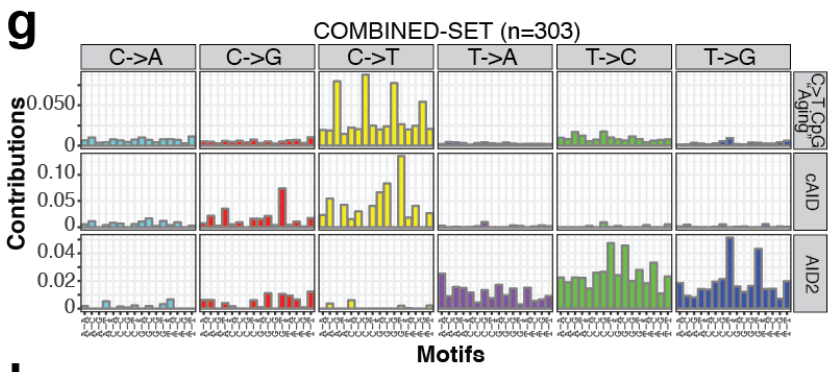


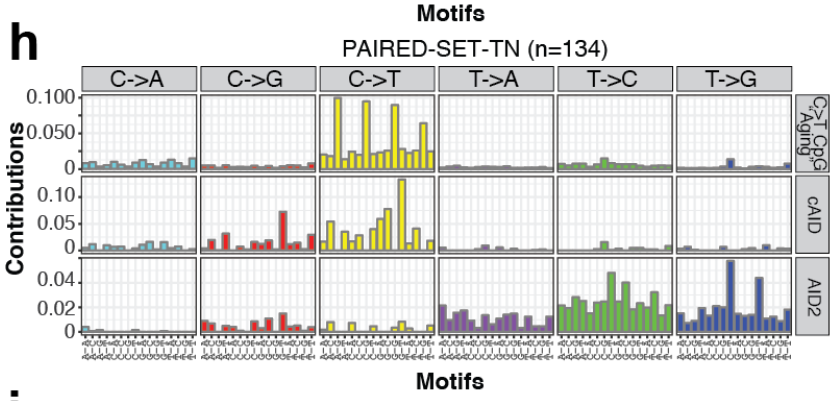
Supplementary Figure 5: Significant mutation clustering at protein structures and additional *BRAF* mutation details. a-d, Crystal structures of CREBBP (**a**, PDB: 4pzt), KRAS (**b**, PDB: 4lv6), MAP2K1 (MEK1, **c**, PDB: 3w8q) and PTPN6 (SHP1, **d**, PDB: 4grz) in grey. Mutated residues in red and color intensity scales with number of mutations. Polar interactions in yellow. Ligands in blue (**a**, S-Co-enzyme A; **b**, GDP; **c**, ATPγS; **d**, phosphate). **e,** *BRAF* mutations are shown in the context of the functional domains of BRAF. Analysis reveals clustering of mutations in the P-loop (orange) and activation-loop (cyan) of the kinase domain. Structural and functional consequences for several of these *BRAF* mutations have been analyzed previously[10-12]. Mutations that either activate the kinase domain by abolishing a hydrophobic interaction between P- and activation-loop (green) or result in a reduced kinase activity (red) are noted[10-12]. Since kinase-death mutations transactivate RAF1, the downstream consequences of all mutations are identical - increased phosphorylation and signaling through ERK[10-12].

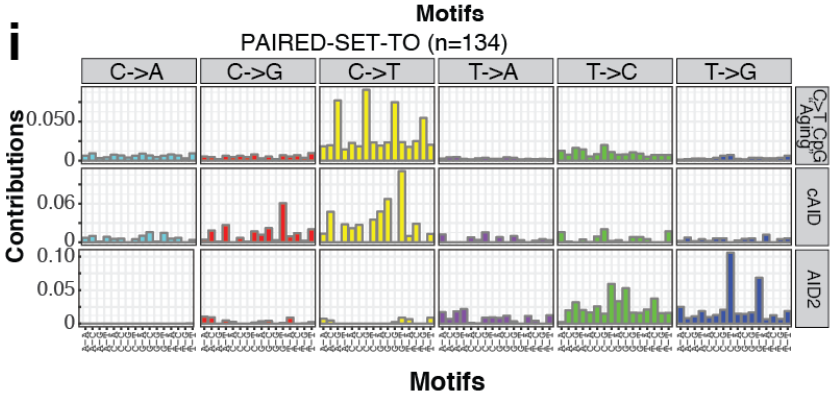


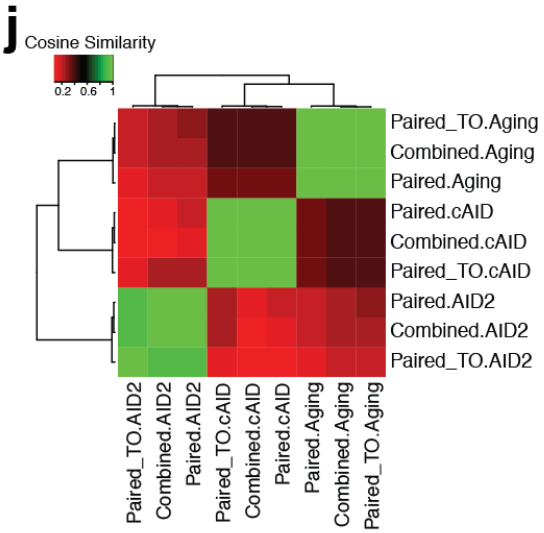
Supplementary Figure 6: Significant mutation clustering at protein interfaces. a, Co-crystal structure of RHOA (grey) and ARHGEF18 (cyan, PDB:4D0N) as a representative example of clustered mutations at the interphase of RHOA to its ARHGEFs (left panel, side view; right panel, 90° rotation around vertical axis of left panel). Mutated residues are labeled in black and shown in red and color intensity scales with number of mutations. Previous studies also described *RHOA* mutations that perturb interactions with ARHGEFs in other tumors[13,14]. **b,** Crystal structure of RHOA (grey, PDB ID: 1dpf). Of note, mutations in RHOA do not affect the catalytic pocket surrounding GDP (blue); instead, the mutations perturb the interphase with ARHGEFs (highlighted in yellow; Supplementary Table 3c, list of all *CLUMPS* at interfaces/*EMPRINT* results). **c,** Model of $RHOA^{wt}$ (top) and $RHOA^{mut}$ (bottom) function. ARHGEFs serve as guanosine exchange factors (GEF) facilitating the replacement of GDP (blue) by GTP (red). Active $RHOA^{wt}$-GTP blocks migration and Pi3K/AKT signaling. Mutations in RHOA ($RHOA^{mut}$) prevent the binding of ARHGEFs, keeping $RHOA^{mut}$ in its inactive GDP state and preventing negative regulation of migration and Pi3K/AKT signaling. **d,** Co-crystal structure FBXW7 (grey) and cyclin E1 (CCNE1, blue). Mutations in FBXW7 at the interphase to the CCNE1 degron are labeled in black. **e,** Model of $FBXW7^{mut}$ function. The $SCF^{FBXW7\text{-}wt}$ (complex of SKP1, CUL1 and $FBXW7^{wt}$) recognizes and targets cyclin E1 (CCNE1) for proteasomal degradation by ubiquitination (Ub)[15]. Mutations in FBXW7 ($FBXW7^{mut}$) perturb the recognition of cyclin E1 and its subsequent proteasomal degradation.

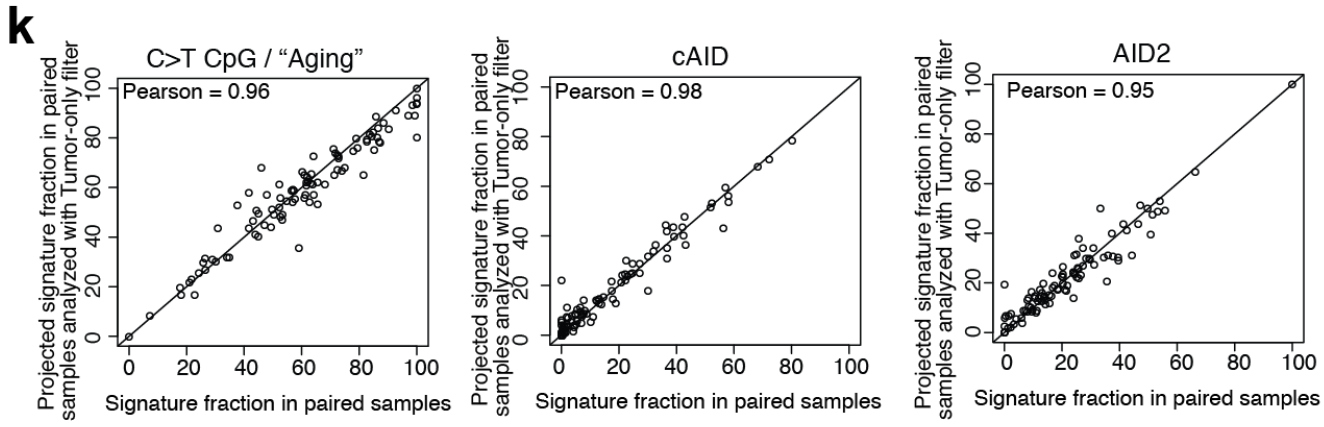

a
C->A
C->G
C->T
T->A
T->C
T->G
Contributions
SNVs
304-A MSI
304-B
304-C
Tri-nucleotide sequence motifs
b
Clustered mutations
Non-clustered mutations
Frequency
log10 (NMD)
c
C>T CpG "Aging"
cAID
AID2
-log10 (observed p value)
-log10(expected p value)
d
Pearson = 0.87
p = 0.01
Pearson = 0.24
p = 0.61
Pearson = 0.33
p = 0.47
Signature activity
Age
e
2p11.1/Igκ
6p21.2/PIM1
14q32.33/IgH
18q21.33/BCL2
22q11.22/Igλ
NMD
Chromosome
f
C>T CpG/"Aging"
cAID
AID2
Signature proportion
Mutations


g
COMBINED-SET (n=303)
C->A
C->G
C->T
T->A
T->C
T->G
C>T CpG Aging
cAID
AID2
Contributions
Motifs

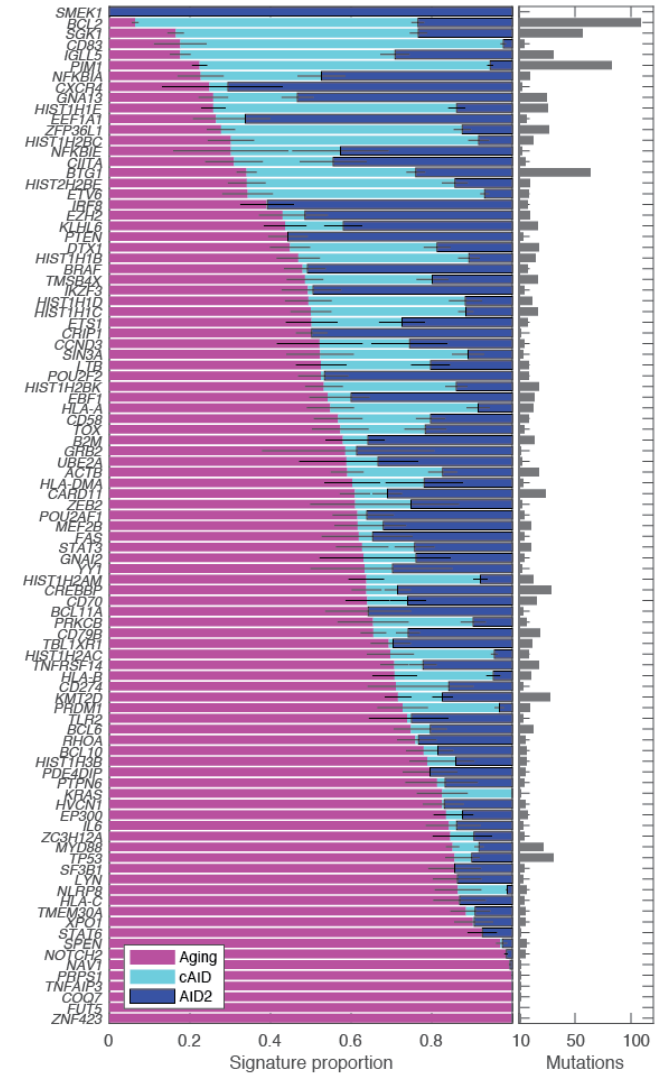

h
PAIRED-SET-TN (n=134)
C->A
C->G
C->T
T->A
T->C
T->G
C>T CpG Aging
cAID
AID2
Contributions
Motifs


i
PAIRED-SET-TO (n=134)
C->A
C->G
C->T
T->A
T->C
T->G
C>T CpG Aging
cAID
AID2
Contributions
Motifs

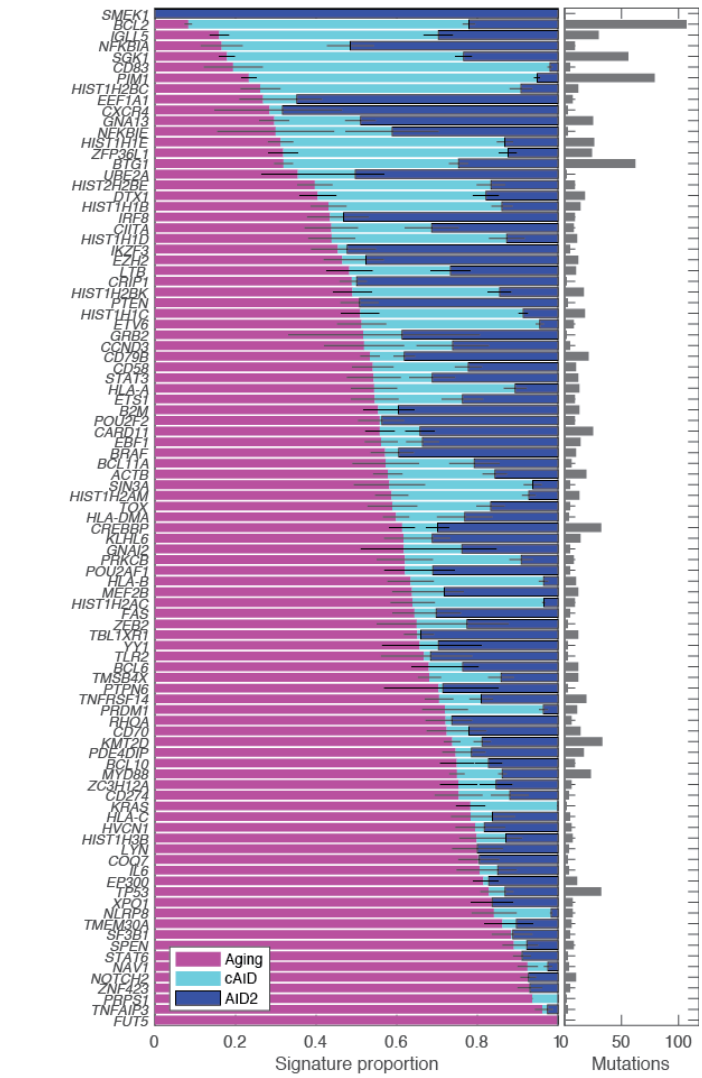

j
Cosine Similarity
0.2
0.6
1
Paired_TO.Aging
Combined.Aging
Paired.Aging
Paired.cAID
Combined.cAID
Paired_TO.cAID
Paired.AID2
Combined.AID2
Paired_TO.AID2

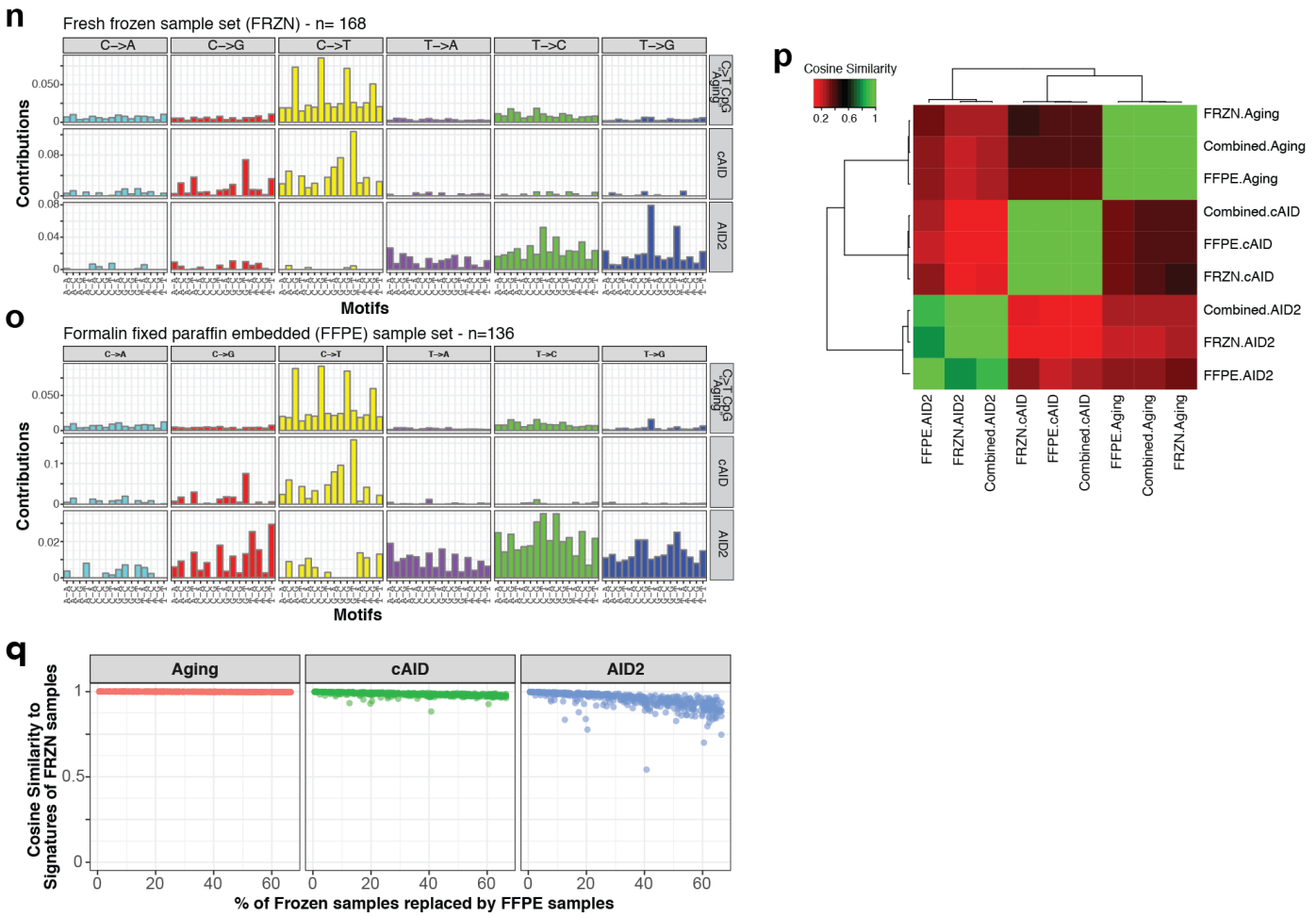

k
C>T CpG / "Aging"
Pearson = 0.96
cAID
Pearson = 0.98
AID2
Pearson = 0.95
Projected signature fraction in paired samples analyzed with Tumor-only filter
Signature fraction in paired samples

l


PAIRED-SET-TN
SMEK1
BCL2
SGK1
CD83
IGLL5
PIM1
NFKBIA
CXCR4
GNA13
HIST1H1E
EEF1A1
ZFP36L1
HIST1H2BC
NFKBIE
CIITA
BTG1
HIST2H2BE
ETV6
IRF8
EZH2
KLHL6
PTEN
DTX1
HIST1H1B
BRAF
TMSB4X
IKZF3
HIST1H1D
HIST1H1C
ETS1
CHIP1
CCND3
SIN3A
LTB
POU2F2
HIST1H2BK
EBF1
HLA-A
CD58
TOX
B2M
GRB2
UBE2A
ACTB
HLA-DMA
CARD11
ZEB2
POU2AF1
MEF2B
FAS
STAT3
GNAI2
YY1
HIST1H2AM
CREBBP
CD70
BCL11A
PRKCB
CD79B
TBL1XR1
HIST1H2AC
TNFRSF14
HLA-B
CD274
KMT2D
PHDM1
TLR2
BCL6
RHOA
BCL10
HIST1H3B
PDE4DIP
PTPN6
KRAS
HVCN1
EP300
IL6
ZC3H12A
MYD88
TP53
SF3B1
LYN
NLRP8
HLA-C
TMEM30A
XPO1
STAT6
SPEN
NOTCH2
NAV1
PRPS1
TNFAIP3
COQ7
FUT5
ZNF423
Aging
cAID
AID2
0
0.2
0.4
0.6
0.8
10
50
100
Signature proportion
Mutations

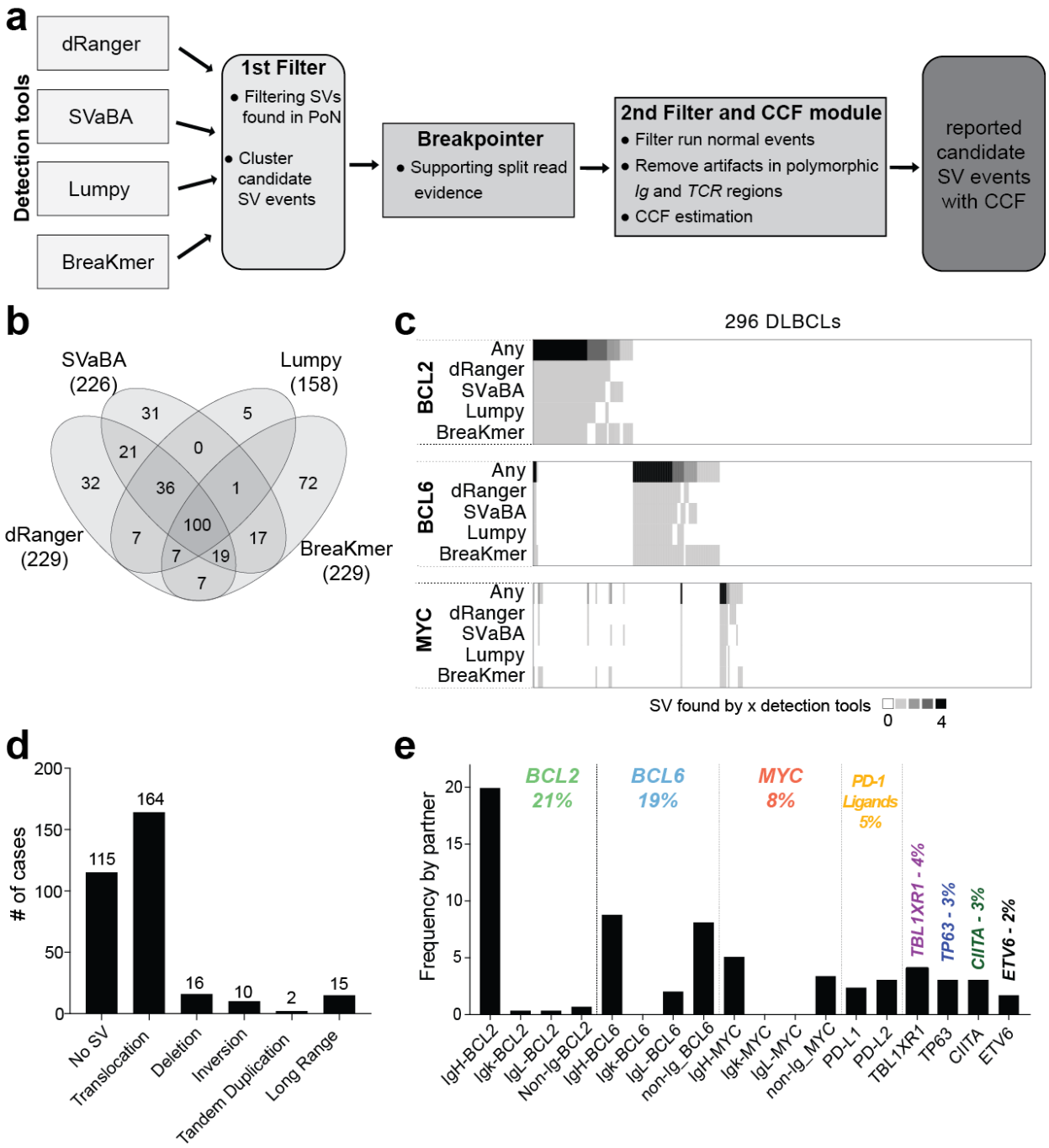
m


PAIRED-SET-TO
SMEK1
BCL2
IGLL5
NFKBIA
SGK1
CD83
PIM1
HIST1H2BC
EEF1A1
CXCR4
GNA13
NFKBIE
HIST1H1E
ZFP36L1
BTG1
UBE2A
HIST2H2BE
DTX1
HIST1H1B
IRF8
CIITA
HIST1H1D
IKZF3
EZH2
LTB
CHIP1
HIST1H2BK
PTEN
HIST1H1C
ETV6
GRB2
CCND3
CD79B
CD58
STAT3
HLA-A
ETS1
B2M
POU2F2
CARD11
EBF1
BRAF
BCL11A
ACTB
SIN3A
HIST1H2AM
TOX
HLA-DMA
CREBBP
KLHL6
GNAI2
PRKCB
POU2AF1
HLA-B
MEF2B
HIST1H2AC
FAS
ZEB2
TBL1XR1
YY1
TLR2
BCL6
TMSB4X
PTPN6
TNFRSF14
PRDM1
RHOA
CD70
KMT2D
PDE4DIP
BCL10
MYD88
ZC3H12A
CD274
KRAS
HLA-C
HVCN1
HIST1H3B
LYN
COQ7
IL6
EP300
TP53
XPO1
NLRP8
TMEM30A
SF3B1
SPEN
STAT6
NAV1
NOTCH2
ZNF423
PRPS1
TNFAIP3
FUT5
Aging
cAID
AID2
0
0.2
0.4
0.6
0.8
10
50
100
Signature proportion
Mutations

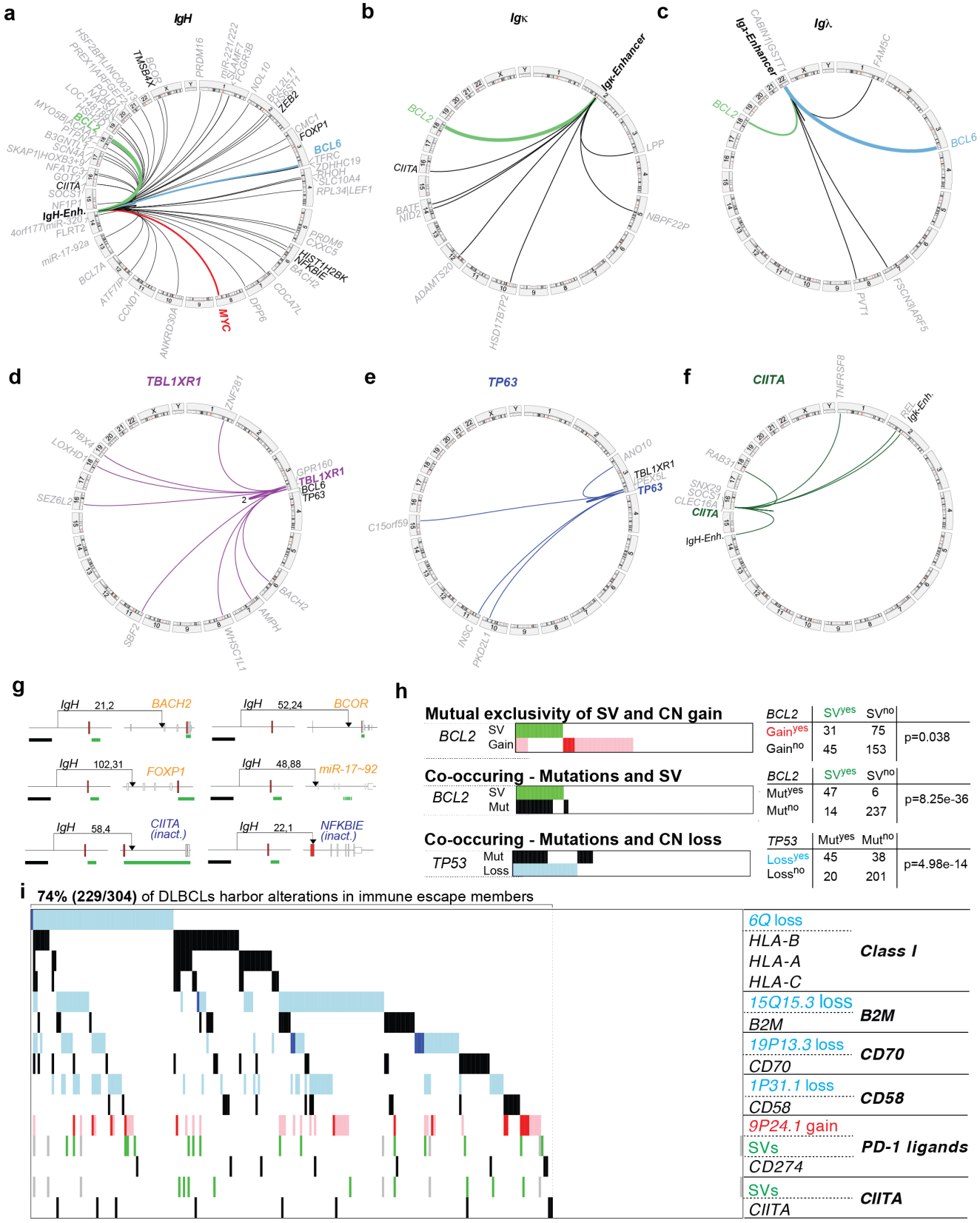


Supplementary Figure 7. Supporting data for mutational signature analysis. a, *De novo* signature extraction for 304 DLBCL samples identified a putative microsatellite instability (MSI) signature in addition to two signatures, 304-B and 304-C. **b,** Based on the bimodal distribution of nearest mutational distance (NMD), all SNV mutations were partitioned into two groups of clustered (NMD ≤ 1kb) and non-clustered mutations (NMD > 1kb). **c,** The Q-Q plots for the gene-level signature enrichment analysis in the remaining 303 DLBCLs after removing the MSI case (see Methods for details). **d,** Correlations of signature activity to the age at diagnosis across seven age groups (n=303 DLBCLs). Box plots represent a distribution of each signature activity of samples belonging to each age bin (line, median; box, interquartile range [IQR]; whiskers,1.5x IQR). The Pearson correlation was calculated between the median signature activity and the median age in each age group. **e,** Rainfall plots of all mutations by mutational signature. Vertical axis illustrates the NMD, horizontal axis the genomic location. Ig loci as loci of physiologic hypermutation are highlighted in blue, *6p21.2/PIM1* and *18q21.33/BCL2* as loci of aberrant somatic hypermutation are visualized in pink. Clustered mutations (NMD ≤ 1kb) below the dotted red line. **f,** For all significantly mutated genes, the relative contribution of each mutational process is visualized (C>T CpG/"Aging", purple; cAID, cyan; AID2, blue). Genes were ordered from top to bottom by the fraction of aging signature. Histogram to the right reports the number of mutations. Error bars show the standard error of the mean. **g-i,** Normalized signature profiles determined by *de-novo* signature extraction for the combined sample set (COMBINED-SET, n=303; **g**), the paired sample set analyzed with the matched normal samples (PAIRED-SET-TN, n=134; **h**), and the paired sample set analyzed without the matched normal (PAIRED-SET-TO, n=134, **i**); **j,** Heatmap of the cosine similarity of three signatures among COMBINE-SET, PAIRED-SET-TN, and PAIRED-SET-TO; **k,** Gene-level signature fraction of the C>T CpG/Aging signature (left), cAID signature (middle), and AID2 signature (right)

across CCGs (n=>10mutations) between PAIRED-SET-TN (n=134; x-axis) and PAIRED-SET-TO (n=134; y-axis). Note that the activity of PAIRED-SET-TO was determined by the projection onto the signature profiles of PAIRED-SET-TN. **l, m,** Signature fractions for CCGs by using **(l)** or not using **(m)** the patient-matched normal sample. Error bars show the standard error of the mean. See f for details. **n,o,** Normalized signature profiles determined by *de-novo* signature extraction for fresh frozen samples (FRZN-SET, n=168) (**n**) and for FFPE samples (FFPR-SET, n=136) (**o**). **p,** Heatmap of the cosine similarity of three signatures among FRZN-SET, FFPE-SET, and COMBINE-SET. **q,** Cosine similarity of Aging (left), cAID (middle), and AID2 (right) signature extracted for 500 pooled sample sets as a function of a fraction of FFPE samples. In each experiment, randomly chosen fresh-frozen samples were replaced by the same number of random FFPE samples.

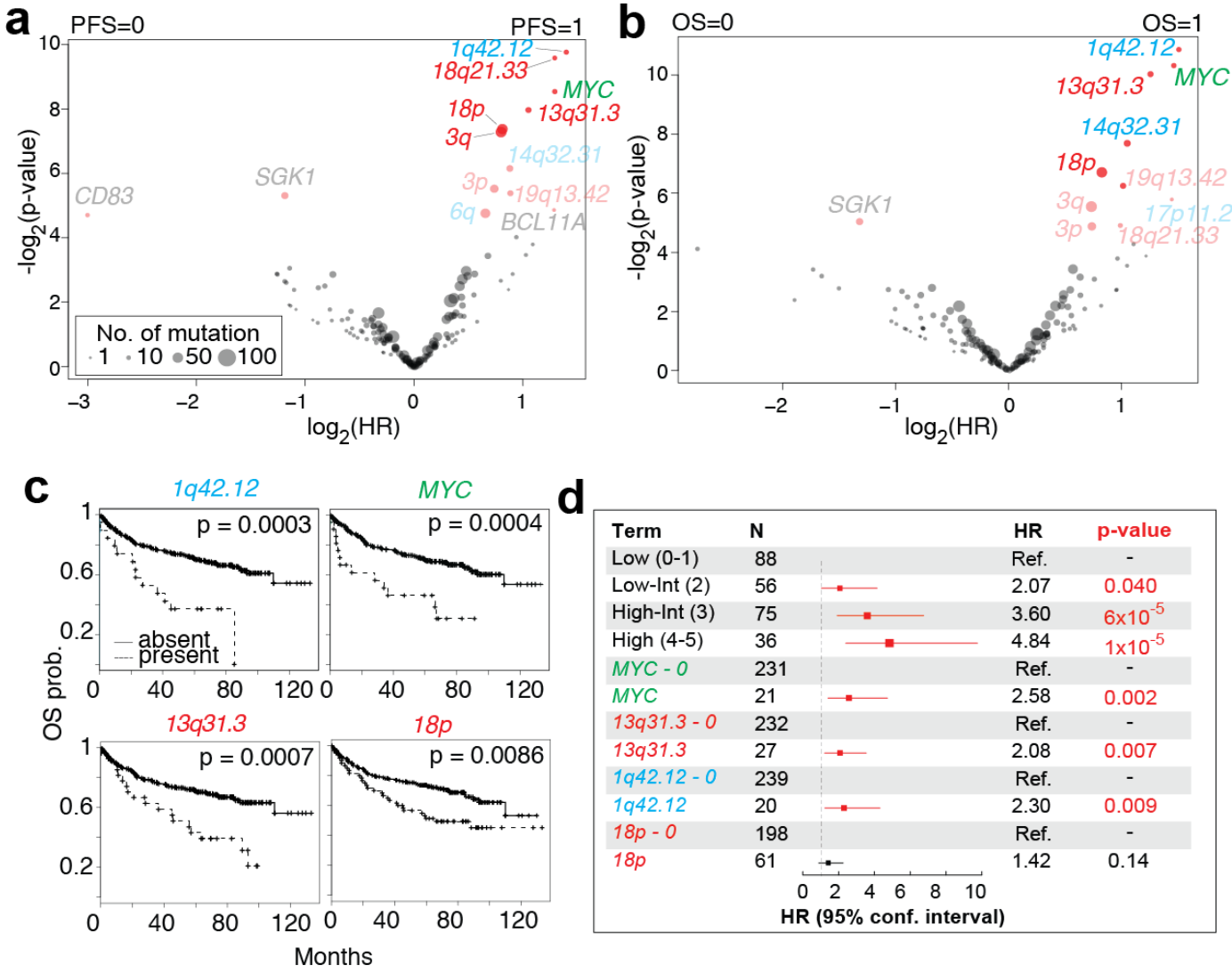


Supplementary Figure 8. Chromosomal rearrangements – pipeline and summary statistics. a, Schematic overview of the analytical pipeline to detect SVs and their CCF. The outputs of four different detectors, dRanger[16], *SVaBA*, Lumpy[17] and BreaKmer[18], were clustered and inputted into Breakpointer[16] to obtain supporting split read evidence and a unified count read of the reference and alternate allele. SVs found with less than 4 total reads, SVs found in a Panel of Normals (PoNs), SV that were part of polymorphic *Ig* and *TCR* regions and artifacts in manual review were filtered out. For the remainder of events, the CCF was calculated as described in the Methods. **b,** Venn Diagram visualizing the overlap of SVs identified by each detector. **c,** Heatmap illustrating the detector evidence for chromosomal rearrangements involving *BCL2, BCL6* and *MYC*. **d,** Summary of SV types in all samples with available SV data (n=296). Of note, translocations of *BCL2* and *BCL6* were largely mutual exclusive (one-sided Fisher's exact test, $p=8.6x10^{-4}$). **e,** Summary of the most frequent SVs ranked by frequency.

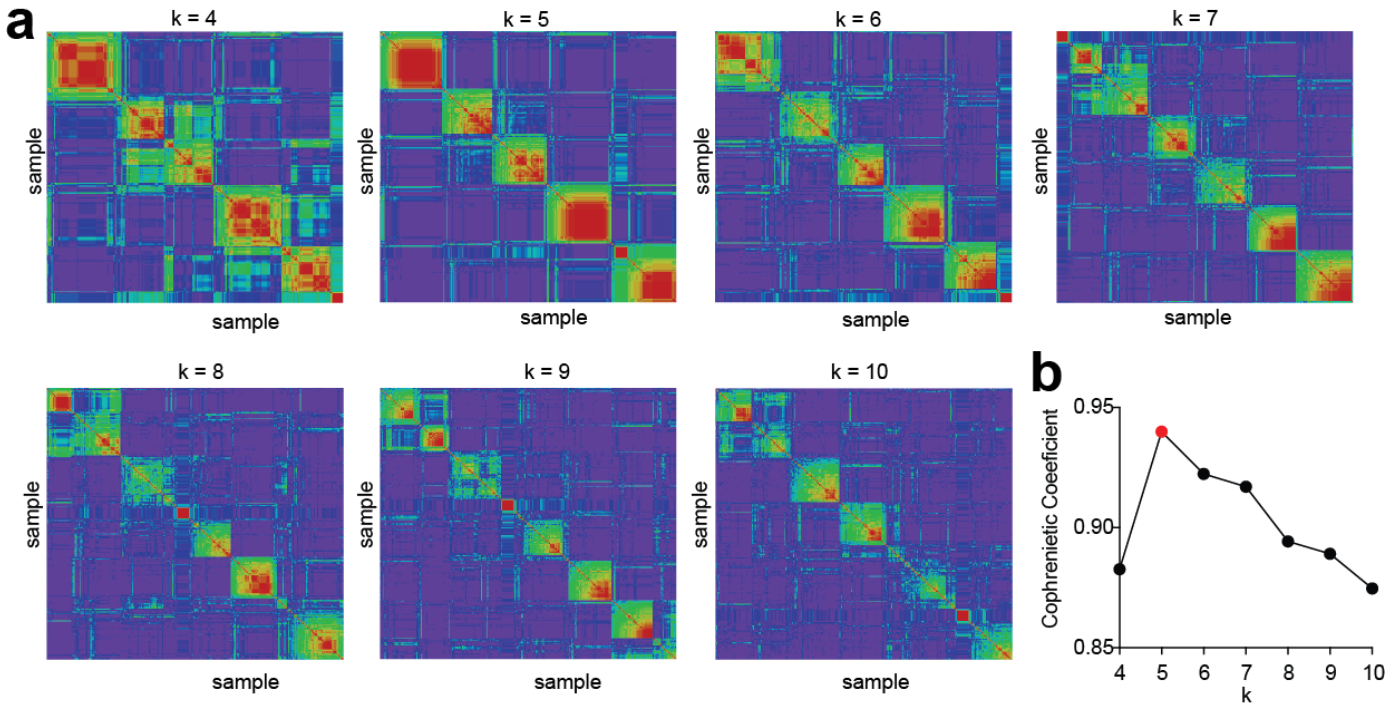


Supplementary Figure 9. Additional chromosomal rearrangements and mutual exclusivity/co-occurrence visualization. a-f, Circos plots of all detected chromosomal rearrangements involving the *IgH* (a), *Igk* (b), *Igl* (c), *TBL1XR1* (d), *TP63* (e) and *CIITA* (f) loci. Line thickness correlates to number of events. Partner genes in grey, if significantly mutated in black. For consistency, *BCL2, BCL6* and *MYC* are in green,

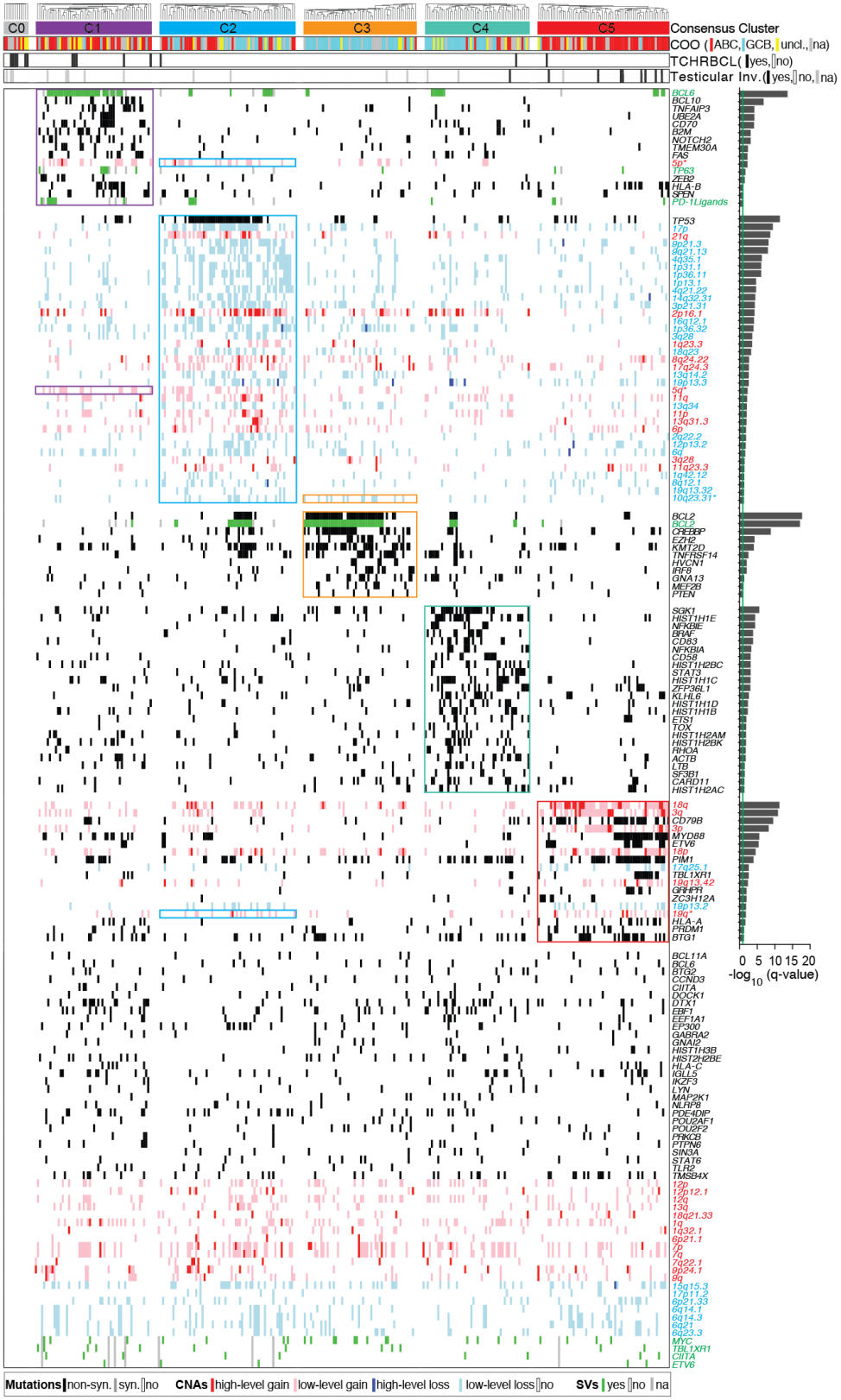
blue and red, as in Fig. 3 of the main manuscript. **g,** Selected translocations between *IgH* and partner genes are plotted in their genomic context. Breakpoints are visualized by an arrow; the two numbers on the arrow indicate split read count followed by read pair count supporting this chromosomal rearrangement. Boxes indicate exons; red box, indicates first coding exon; ORF in green; enhancer element in black. Translocations are activating (orange partner gene) or inactivating by destroying the ORF (dark blue). All diagrams display the *IgH* partner in the coding direction. **h,** Color-coded matrices that visualize significant mutual exclusivity of SVs and CN gains in *BCL2* (top, p=0.038), co-occurring of SV and mutations in *BCL2* (middle, $p=8.25e^{-36}$) and co-occurring of single CN loss and mutations in *TP53* (bottom, $p=4.98e^{-14}$). Mutations, black; SV, green; single CN loss, cyan; low grade CN gain, pink; high grade CN gain, red. Contingency table of events in the full cohort (n=304) and p-value obtained by a one-sided Fisher's Exact test are displayed to the right. **i,** Color-coded matrix shows genetic alterations in indicated immune evasion molecules. See for color code legend of main **Fig. 4**.



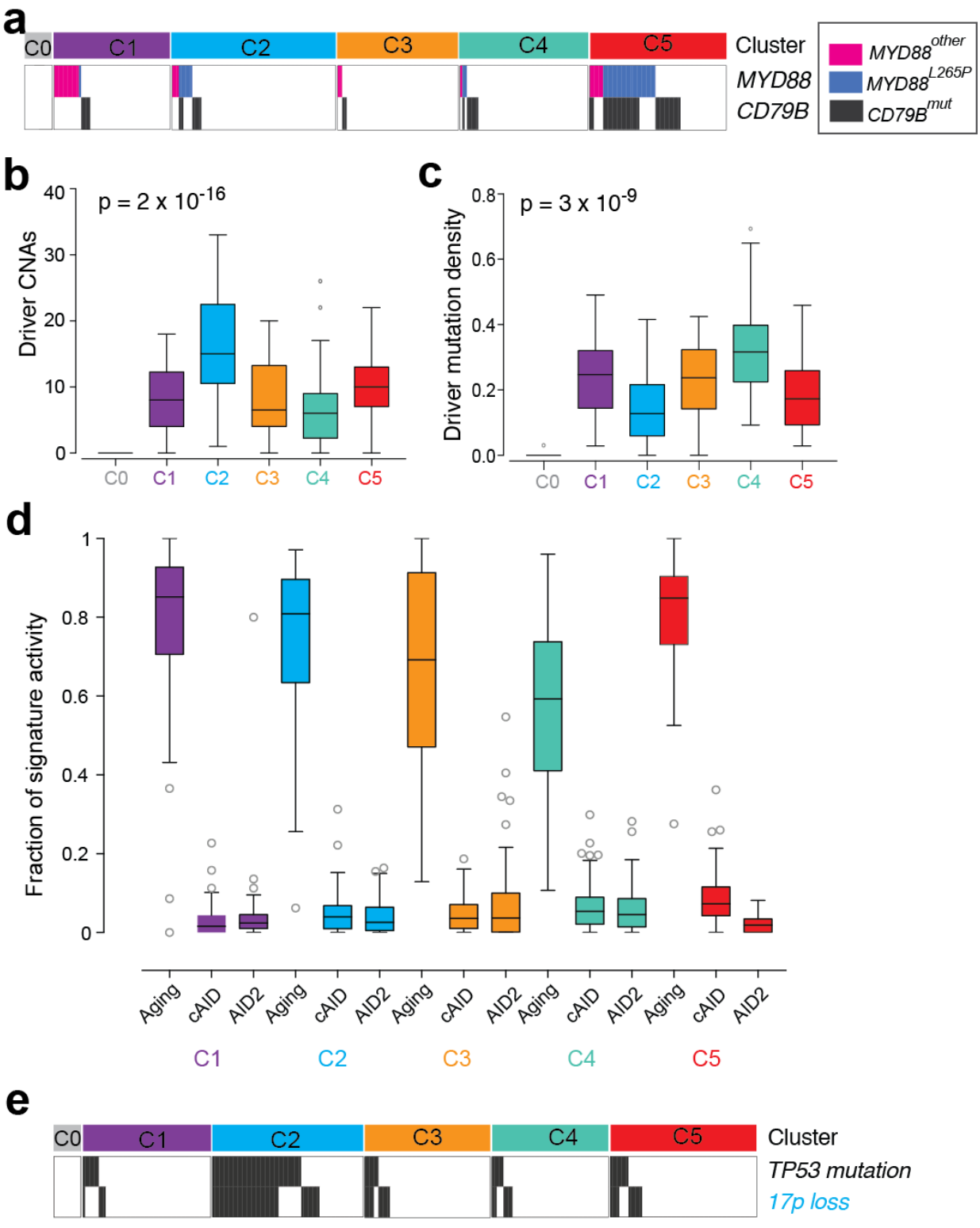
Supplementary Figure 10. Supporting outcome analyses of individual genetic alterations. a,b, Individual genetic features detected in ≥ 3% of the R-CHOP cohort (150 genetic drivers) were assessed for their association to PFS (**a**) and OS (**b**) in univariate Cox regression models (q-value <0.2). Volcano plots show hazard ratio (x-axis) vs. significance (y-axis). Size of dots represent the number of mutations; the color of dots represent significance (p-value<0.05, light red; p-value>0.05, grey). Alterations with significant p-values are labeled (q-value<0.2: CN gains, red; CN loss, blue; SVs, green; q-value>0.2: Mutations, grey; CN gain, pale light red; CN blue: light blue). **c,** Kaplan Meier plots for significant factors in univariate model predicting OS in the R-CHOP treated cohort with OS data (n= 259) that were also independent to each other in a multivariate model; alterations absent, solid line; alterations present, dashed line; p-values derived from log-rank test. **d,** Forest plots visualize the multivariate analysis of IPI risk groups and individual genetic factors for OS in the R-CHOP treated cohort with OS data (n= 259).

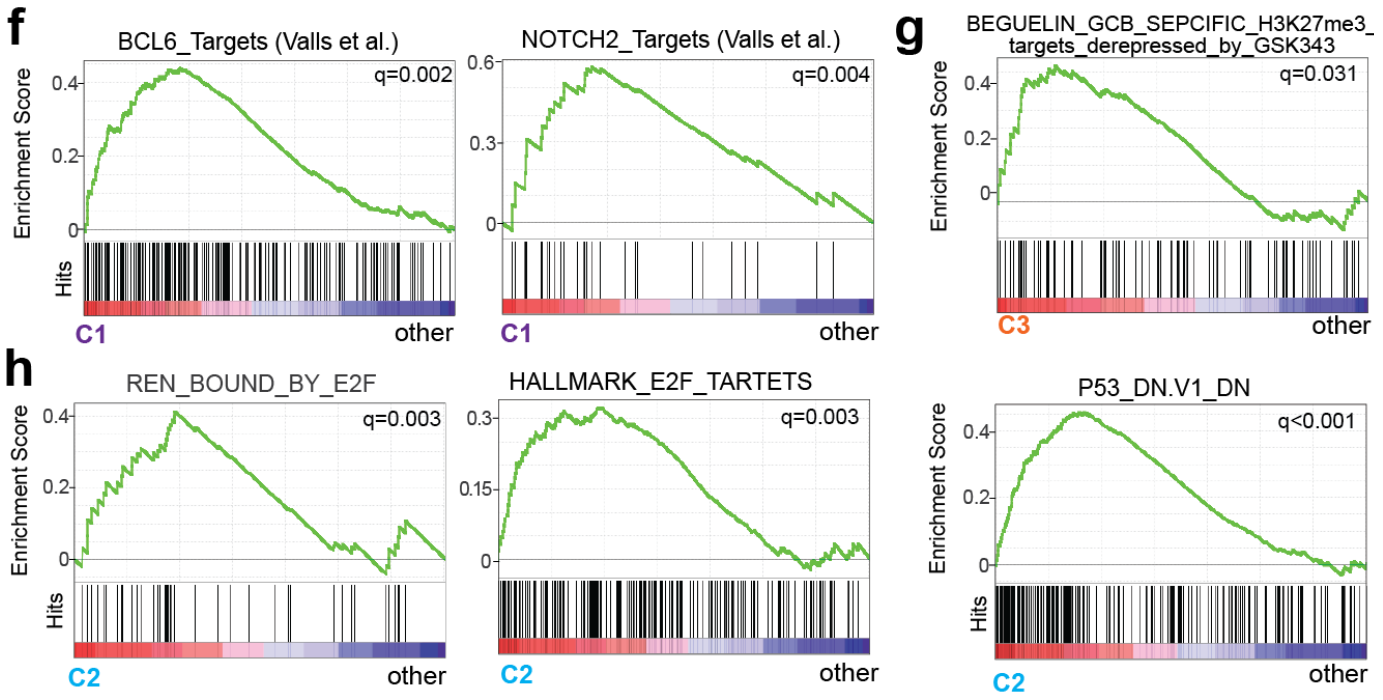


Supplementary Figure 11. Consensus clustering. a, Consensus plots for k=4 to k=10 cluster solutions. **b,** Cophenetic coefficient for k=4 to k=10 cluster solutions.

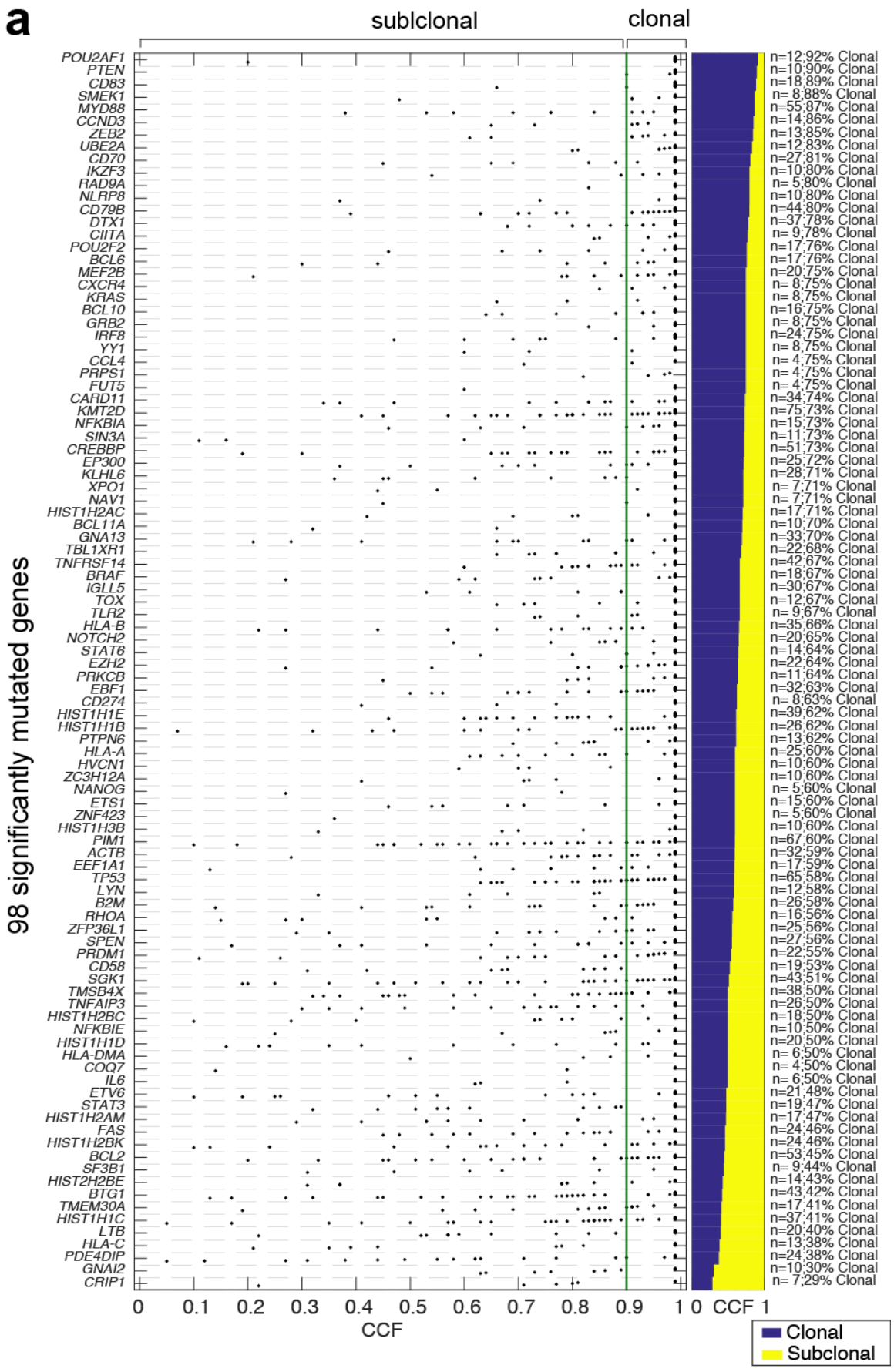


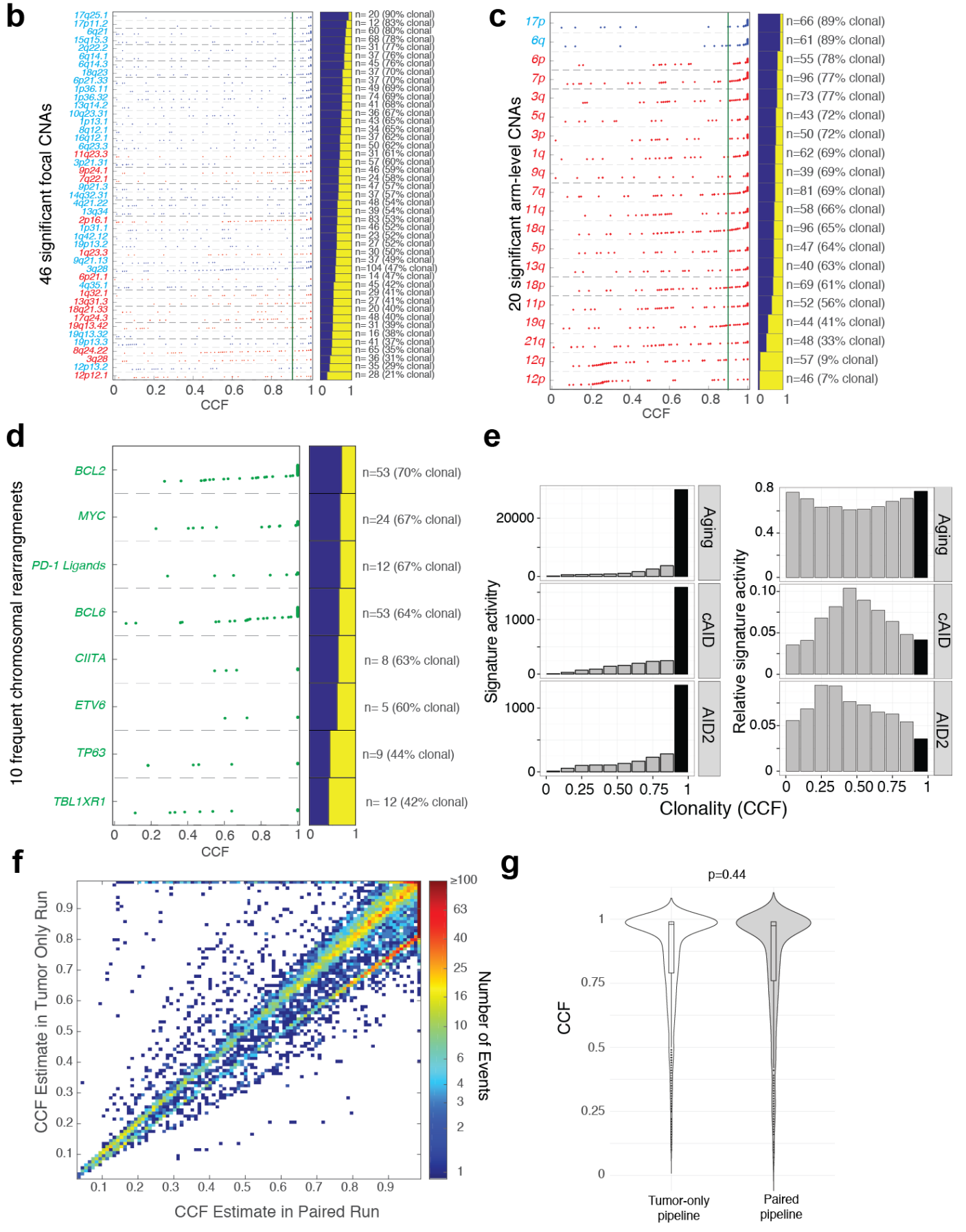
Supplementary Figure 12. Expanded consensus clustering gene-sample matrix. Clusters C1-C5 with their associated landmark genetic alterations are visualized as in Fig. 4 (top). At the bottom, the additional recurrent alterations that were not associated with specific clusters are shown. See also legend of main **Fig. 5.**

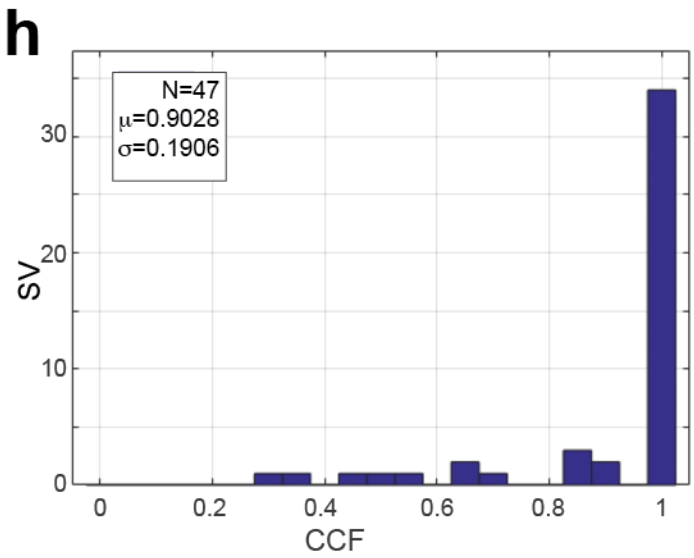

a
C0
C1
C2
C3
C4
C5
Cluster
MYD88
CD79B
MYD88other
MYD88L265P
CD79Bmut
b
p = 2 x 10-16
Driver CNAs
c
p = 3 x 10-9
Driver mutation density
d
Fraction of signature activity
Aging
cAID
AID2
e
TP53 mutation
17p loss



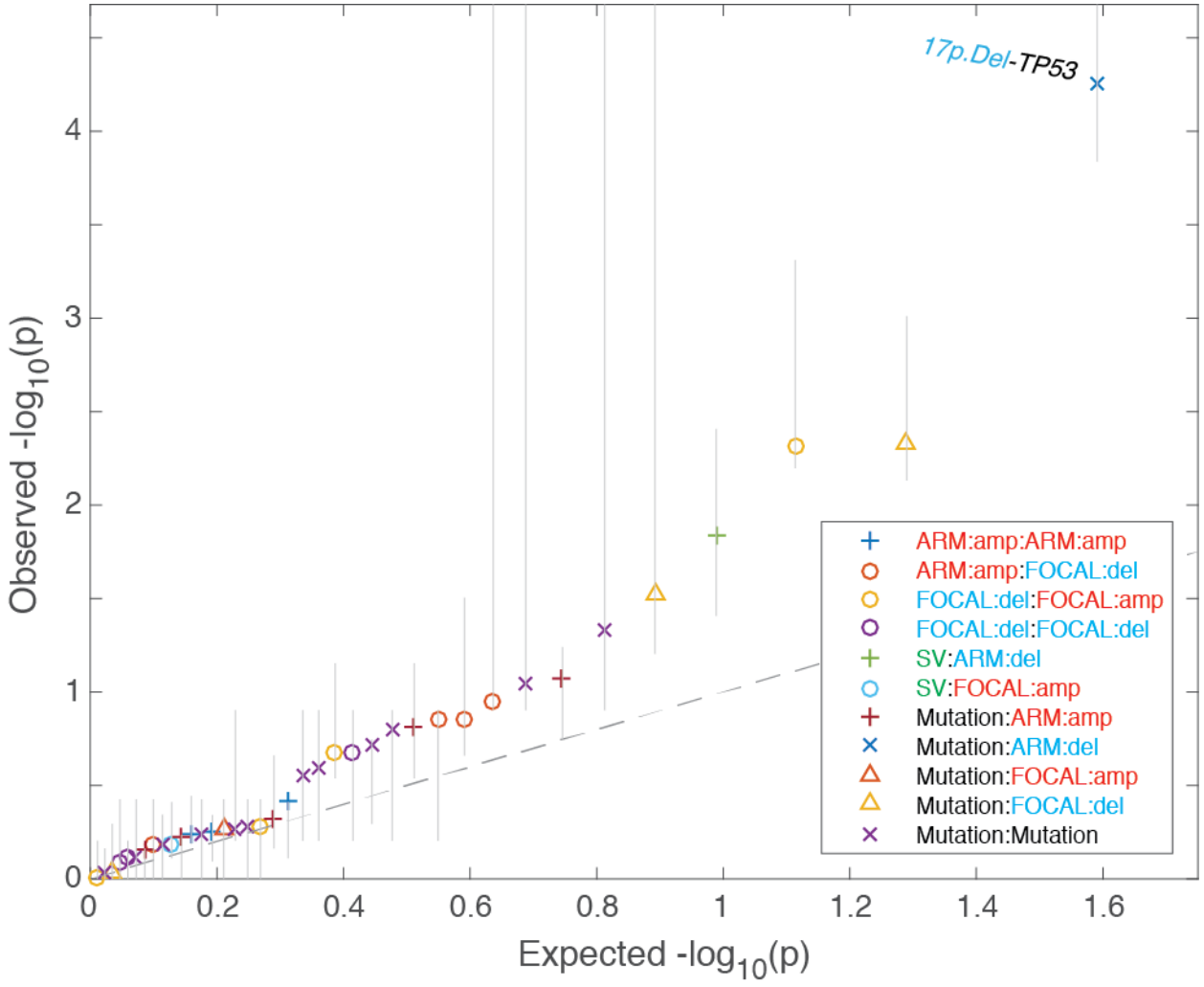
Supplementary Figure 13. Summary of genetic alterations by cluster. a, Incidence of *MYD88* and *CD79B* mutations in clusters C0-C5 (Fig.4, main manuscript). Types of *MYD88* mutations are color-coded (*MYD88*^{L265P}, blue; *MYD88*^{other}, pink). In cluster C5, *MYD88* and *CD79B* mutations are more frequent, MYD88 mutations are more likely to be *MYD88*^{L265P} and *MYD88* and *CD79B* mutations are more likely to be concordant. **b,** Drivers SCNAs in clusters C0-C5 (n=304). The p-values is obtained using a two-sided Mann-Whitney U test. **c,** Mutation density in clusters C0-C5 (n=304). The p-values is obtained using a two-sided Mann-Whitney U test. **d,** Fraction of mutational signature activity for each cluster C1-C5 (n=292). **b-d,** Data visualized as a Tukey box plots (line, median; box, interquartile range [IQR]; whiskers, 1,5x above or below median). **e,** Incidence of *TP53* mutations and 17p loss across clusters C0-C5 (n=304). Bi-allelic inactivation of TP53 in C2 is significantly more frequent than in other clusters (p= $8x10^{-14}$; two-sided Fisher's exact test). **f,** Gene set enrichment analysis (GSEA) plot of BCL6 and NOTCH2 target gene sets[19] in C1 DLBCLs compared to other DLBCLs. **g,** GSEA of a functionally defined EZH2 target gene list[20] in C3 DLBCLs compared to other DLBCLs. **h,** GSEA of E2F and TP53 target genes (MSigDB; http://software.broadinstitute.org/gsea/msigdb;[21]) in C2 DLBCLs compared to the other DLBCLs. **f-h,** The GSEA analyses were performed in all samples with available gene expression data and Cluster C1-C5 annotation (n=131).


a
sublclonal
clonal
98 significantly mutated genes
POU2AF1 n=12;92% Clonal
PTEN n=10;90% Clonal
CD83 n=18;89% Clonal
SMEK1 n= 8;88% Clonal
MYD88 n=55;87% Clonal
CCND3 n=14;86% Clonal
ZEB2 n=13;85% Clonal
UBE2A n=12;83% Clonal
CD70 n=27;81% Clonal
IKZF3 n=10;80% Clonal
RAD9A n= 5;80% Clonal
NLRP8 n=10;80% Clonal
CD79B n=44;80% Clonal
DTX1 n=37;78% Clonal
CIITA n= 9;78% Clonal
POU2F2 n=17;76% Clonal
BCL6 n=17;76% Clonal
MEF2B n=20;75% Clonal
CXCR4 n= 8;75% Clonal
KRAS n= 8;75% Clonal
BCL10 n=16;75% Clonal
GRB2 n= 8;75% Clonal
IRF8 n=24;75% Clonal
YY1 n= 8;75% Clonal
CCL4 n= 4;75% Clonal
PRPS1 n= 4;75% Clonal
FUT5 n= 4;75% Clonal
CARD11 n=34;74% Clonal
KMT2D n=75;73% Clonal
NFKBIA n=15;73% Clonal
SIN3A n=11;73% Clonal
CREBBP n=51;73% Clonal
EP300 n=25;72% Clonal
KLHL6 n=28;71% Clonal
XPO1 n= 7;71% Clonal
NAV1 n= 7;71% Clonal
HIST1H2AC n=17;71% Clonal
BCL11A n=10;70% Clonal
GNA13 n=33;70% Clonal
TBL1XR1 n=22;68% Clonal
TNFRSF14 n=42;67% Clonal
BRAF n=18;67% Clonal
IGLL5 n=30;67% Clonal
TOX n=12;67% Clonal
TLR2 n= 9;67% Clonal
HLA-B n=35;66% Clonal
NOTCH2 n=20;65% Clonal
STAT6 n=14;64% Clonal
EZH2 n=22;64% Clonal
PRKCB n=11;64% Clonal
EBF1 n=32;63% Clonal
CD274 n= 8;63% Clonal
HIST1H1E n=39;62% Clonal
HIST1H1B n=26;62% Clonal
PTPN6 n=13;62% Clonal
HLA-A n=25;60% Clonal
HVCN1 n=10;60% Clonal
ZC3H12A n=10;60% Clonal
NANOG n= 5;60% Clonal
ETS1 n=15;60% Clonal
ZNF423 n= 5;60% Clonal
HIST1H3B n=10;60% Clonal
PIM1 n=67;60% Clonal
ACTB n=32;59% Clonal
EEF1A1 n=17;59% Clonal
TP53 n=65;58% Clonal
LYN n=12;58% Clonal
B2M n=26;58% Clonal
RHOA n=16;56% Clonal
ZFP36L1 n=25;56% Clonal
SPEN n=27;56% Clonal
PRDM1 n=22;55% Clonal
CD58 n=19;53% Clonal
SGK1 n=43;51% Clonal
TMSB4X n=38;50% Clonal
TNFAIP3 n=26;50% Clonal
HIST1H2BC n=18;50% Clonal
NFKBIE n=10;50% Clonal
HIST1H1D n=20;50% Clonal
HLA-DMA n= 6;50% Clonal
COQ7 n= 4;50% Clonal
IL6 n= 6;50% Clonal
ETV6 n=21;48% Clonal
STAT3 n=19;47% Clonal
HIST1H2AM n=17;47% Clonal
FAS n=24;46% Clonal
HIST1H2BK n=24;46% Clonal
BCL2 n=53;45% Clonal
SF3B1 n= 9;44% Clonal
HIST2H2BE n=14;43% Clonal
BTG1 n=43;42% Clonal
TMEM30A n=17;41% Clonal
HIST1H1C n=37;41% Clonal
LTB n=20;40% Clonal
HLA-C n=13;38% Clonal
PDE4DIP n=24;38% Clonal
GNAI2 n=10;30% Clonal
CRIP1 n= 7;29% Clonal
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
CCF
0 CCF 1
Clonal
Subclonal

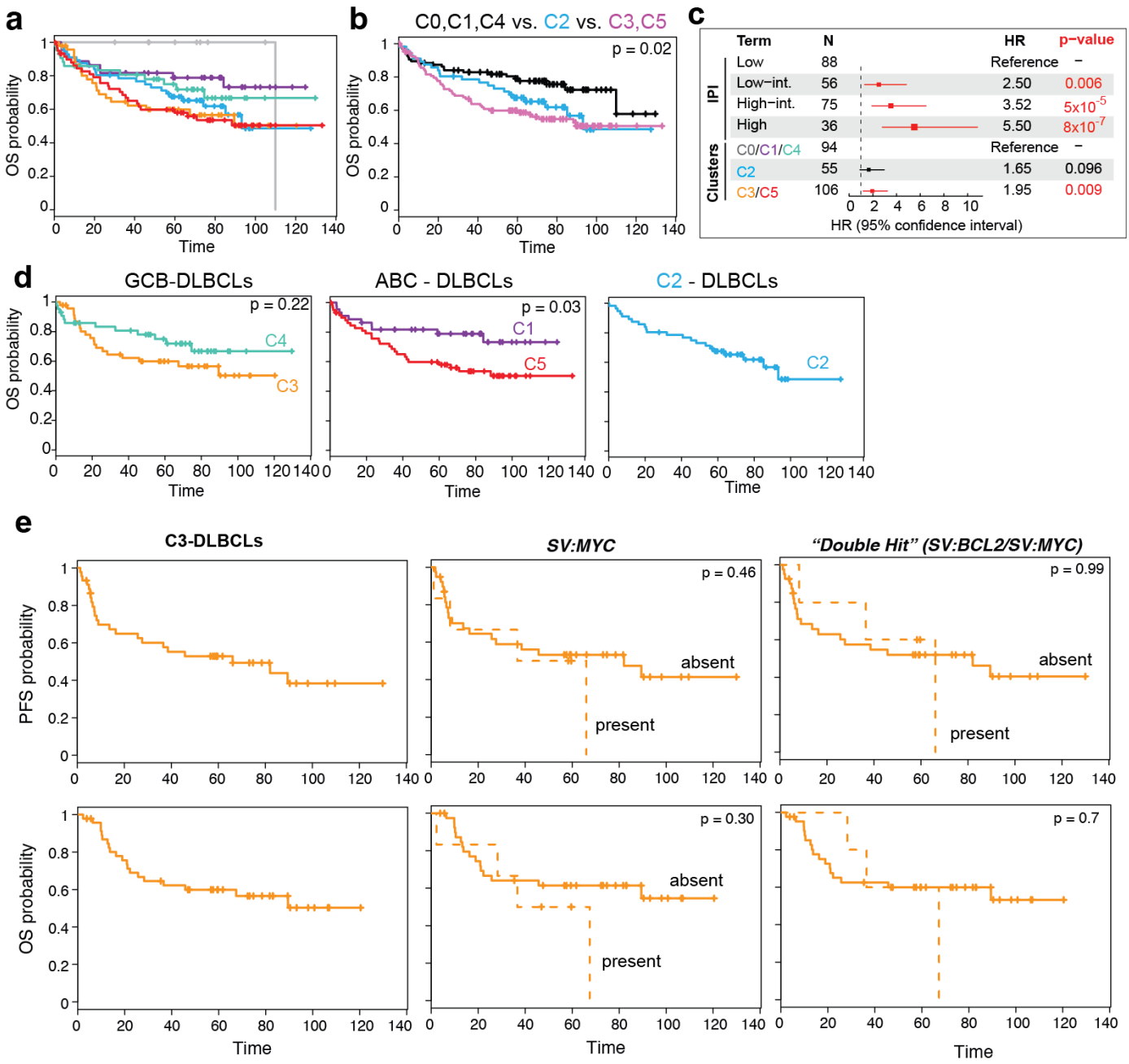

b
46 significant focal CNAs
CCF
c
20 significant arm-level CNAs
CCF
d
10 frequent chromosomal rearrangmenets
BCL2
MYC
PD-1 Ligands
BCL6
CIITA
ETV6
TP63
TBL1XR1
n=53 (70% clonal)
n=24 (67% clonal)
n=12 (67% clonal)
n=53 (64% clonal)
n= 8 (63% clonal)
n= 5 (60% clonal)
n=9 (44% clonal)
n= 12 (42% clonal)
CCF
e
Signature activity
Relative signature activity
Aging
cAID
AID2
Clonality (CCF)
f
CCF Estimate in Tumor Only Run
CCF Estimate in Paired Run
Number of Events
g
p=0.44
CCF
Tumor-only pipeline
Paired pipeline



Supplementary Figure 14. CCF distribution plots for each alteration type and mutational signatures. **a-d,** CCF distribution plots for mutations (a), focal SCNAs (b), arm-level SCNAs (c), chromosomal rearrangements (d). Individual events as scatter plots to the left, Histogram to the right summarizes the overall clonality (blue, clonal; yellow, subclonal). Alterations with a CCF $\geq$ 0.9 are defined as "clonal". **e,** CCF distribution of mutations for each mutational signature. Black, clonal; grey, subclonal. Plotted is the total number of mutations (left) and the fraction of mutations for a given CCF. **f,** Scatter plot of CCF estimates for shared mutations between 134 samples with patient-matched normal samples analyzed either with the tumor-only or tumor-normal pipeline, highlighting for all samples the tight correlation between the CCF called in each pipeline. The main deviation was in the hypermutator case, represented by a secondary diagonal line below the primary one, where the purity was called differently in ABSOLUTE between the paired (0.81), and tumor-only pipeline (0.98). **g,** Violin plot (box, interquartile range [IQR]; line in box, median; whiskers, 1.5x IQR; width of violin plot, density of events) comparing the CCF distributions of CCGs with more than 10 mutations between all tumor-only samples (n=169) and all paired samples (n=135). The p-value is obtained from a two-sided Fisher's exact test comparing clonal and subclonal mutations between the two groups. **h,** Histogram of CCFs for all 47 SVs detected in DLBCL cell lines involving *BCL2, BCL6* or *MYC*. Note that the majority of CCFs indicate clonal alterations, providing an orthogonal evidence that the CCF SV pipeline accurately determines the CCF of likely clonal drivers.



Supplementary Figure 15. Q-Q plot of timing analysis of genetic alterations. For every pair of cancer gene somatic variants (mutations, SVs, SCNAs) in the full DLBCL cohort (n=304), we selected the pairs with at least 4 samples in which the two variants co-occurred with one as clonal and the other sub-clonal. In the absence of event ordering, we would expect these events to have a random ordering according to a two-sided binomial test with $p=0.5^{22}$. The q-q plot of the expected vs. observed $-\log_{10}(p)$ of paired events in the full cohort shows that the bulk of variant pairs have no significant ordering.



Supplementary Figure 16. Association of coordinate genetic signatures with overall survival (OS). a, Kaplan Meier (KM) plot for OS for all clusters, C0 (grey), C1 (purple), C2 (cyan), C3 (orange), C4 (turquoise), C5 (red) to the left. **b,** KM plot for OS for favorable DLBCL clusters (C0, C1,C4) in black, C2-DLBCLs in blue and unfavorable DLBCLs (C3, C5) in pink. The p-value obtained using the log-rank test. **c,** Forest plots visualize HR and p-values obtained from the multivariate analysis of OS including clusters and IPI. **d,** KM plot for OS for the two genetically distinct GCB-DLBCLs (left), ABC-DLBCLs (middle) and C2 DLBCLs. The p-value obtained using the log-rank test. **a-d,** All analyses were performed in the R-CHOP treated cohort with OS information (n=259). **e,** KM plot for PFS (upper row; n=45) and OS (lower row; n=47) for all C3 DLBCLs (left), C3 DLBCLs split by presence/absence of *SV-MYC* (middle) and C3 DLBCLs split by presence/absence of concurrent *SV-MYC* and *SV-BCL2 (*right). The p-value obtained using the log-rank test.

Supplementary Note

Additional Quality Control. The amount of cross-individual contamination for tumor-normal pairs was estimated by *ContEst*[23] and the median contamination was 0.3% (interquartile-range 0.1-0.4%) . Due to the lack of methods for estimating cross-individual contamination without a paired normal, tumor-only samples were assumed to have 2% contamination, which is higher than 98% of the contamination estimates determined by *ContEst* in the paired samples.

A total of 47 samples were omitted due to quality control concerns, including 12 that failed sequencing (no BAM file), 23 samples had coverage that was too low to analyze (less than 75% of exome was callable), 1 was removed because of a high *ContEst* value (greater than 5%) and 11 were removed due to pervasive realignment artifact. Of the remaining 304 samples, one sample was treated as a tumor-only because its paired normal showed evidence of high tumor in normal contamination and two samples where the tumor and normal samples were mixed up.

Estimation of and correction for tumor in normal content (deTiN). As described previously[22], tumor in normal contamination significantly decreases the ability to detect somatic mutations. Therefore, we estimated the presence of tumor contamination in matched normal samples when available using the *deTiN* algorithm[22,24]. Briefly, *deTiN* uses candidate somatic nucleotide variants and allelic copy number events to infer the fraction of tumor cells in a matched normal sample. This estimate was then used to recover somatic events that otherwise would have been rejected by *MuTect* or *Indelocator* due to low-allele fraction presence in the normal.

Germline Somatic Log Odds Filter for Tumor-only Samples. For each event that passed all preceding filters (SNV or Indel), its CCF, purity, ploidy and local copy number were used to determine the log ratio of the probability that its allele fraction is consistent with the allele fraction modeled for a hypothetical germline event and the probability it is consistent with a modeled somatic event. First, the total amount of DNA per cell (in units of copy number) was calculated as follows:

$$D = (1-\alpha)N + (1-C_{CNV})\alpha N + \alpha C_{CNV}(\mu + M)$$

Where α is the sample purity, N is the normal copy number at the site (2 for autosomes, 2 for X in females and 1 for X in males), C_{CNV} is the CCF of any potential SCNAs at that site, and μ and M are the minor and major allele counts of the copy number. Because heterozygous germline sites should always be at 50% allele fraction in the normal component of the tumor, only copy number alterations should affect the predicted allele fraction. Thus, there are two models for germline event allele fraction, depending on whether the germline event is on the minor allele (G_1) or major allele (G_2) of a copy number event in the region (**Supplementary Fig. 2b,c**):

$$G_1 = \frac{(1-\alpha) + (1-C_{CNV})\alpha + \alpha C_{CNV}\mu}{D}$$
$$p_{G_1} = \beta(G_1, n_{ALT}+1, n_{REF}+1)$$

$$G_2 = \frac{(1-\alpha) + (1-C_{CNV})\alpha + \alpha C_{CNV}M}{D}$$
$$p_{G_2} = \beta(G_2, n_{ALT}+1, n_{REF}+1)$$

Where G_1 is the modeled allele fraction for when the germline site event is on the minor allele, and G_2 is the modeled allele fraction for when the germline event is on the major allele. The probability that the observed allele fraction is consistent with this model is calculated based on a beta (β) probability distribution function, where the modeled hypothesis is tested against the actual counts of reference (n_{REF}) and variant (n_{ALT}) reads. For somatic events, 6 separate models must be evaluated depending on the order in which events happen in a tumor (**Supplementary Fig. 2d,e**). Specifically, the models account for the minor and major allele when a somatic event co-occurs with a copy number event (S_1 + S_2), occurs before a copy number event (S_3 + S_4), occurs after a copy number event (S_5), or if it occurs in a different subclone (S_6):

$$S_1 = \frac{\alpha C_{mut} C_{CNV} M}{D}$$
$$p_{S_1} = \beta(S_1, n_{ALT} + 1, n_{REF} + 1)$$

$$S_2 = \frac{\alpha C_{mut} C_{CNV} \mu}{D}$$
$$p_{S_2} = \beta(S_2, n_{ALT} + 1, n_{REF} + 1)$$

$$S_3 = \frac{\alpha C_{mut}(1 - C_{CNV}) + \alpha C_{mut} C_{CNV} M}{D}$$
$$p_{S_3} = \beta(S_3, n_{ALT} + 1, n_{REF} + 1)$$

$$S_4 = \frac{\alpha C_{mut}(1 - C_{CNV}) + \alpha C_{mut} C_{CNV} \mu}{D}$$
$$p_{S_4} = \beta(S_4, n_{ALT} + 1, n_{REF} + 1)$$

$$S_5 = \frac{\alpha C_{mut} C_{CNV}}{D}$$
$$p_{S_5} = \beta(S_5, n_{ALT} + 1, n_{REF} + 1)$$

$$S_6 = \frac{\alpha C_{mut}(1 - C_{CNV})}{D}$$
$$p_{S_6} = \beta(S_6, n_{ALT} + 1, n_{REF} + 1)$$

Where C_{mut} is the CCF of the somatic event as calculated by *ABSOLUTE*. Once the probability that each model is consistent with the data, the log odds ratio of the most likely germline and somatic model, L, becomes the statistic to apply the filter on:

$$p_G = max(p_{G1}, p_{G2})$$
$$p_S = max(p_{S_1}, p_{S_2}, p_{S_3}, p_{S_4}, p_{S_5}, p_{S_6})$$
$$L = \log \frac{p_G}{p_S}$$

Because the allele fraction of a clonal heterozygous somatic event will be similar to a germline heterozygous site at high purity, but falls as the purity goes down, the divergence in L between putative somatic events (events present in the paired analysis) and putative germline events (events present only when paired samples are run without their paired normals) differs greatly depending on purity (**Supplementary Fig. 2f**),

meaning that while for impure samples we can use a very stringent cutoff of 0, for higher purity samples the cutoff must be relaxed to prevent the removal of true somatic events with high clonality.

To calibrate the cutoff, the data consisting of the 147 available non-hypermutator lymphoma samples with paired normal was split into two training sets, each set was split into 10 bins with similar purity and cutoffs were found that preserved 99% of the putative somatic events in each bin and then used to fit a linear model determining the best cutoff depending on purity (**Supplementary Fig. 3a-c**). After applying this filter, the less pure samples of our cohort show nearly the same mutation rate as when run through the pipeline with their paired normal. While this step filters out many of the remaining germline events, no sample reports fewer events after this filter than in the paired pipeline (**Supplementary Fig. 3d-e**).

Clustering and visualization of mutations in protein structures. We overlaid the identified missense mutations found in our cohort onto protein structures from the Protein Data Bank (RCSB PDB; www.rcsb.org)[25] and applied the recently reported *CLUMPS* algorithm[13] to identify significant spatial clustering of mutations in protein structures. Briefly, *CLUMPS* summarizes the pairwise three-dimensional Euclidean distances between mutated residues into a score function and compares the score to a null model obtained by randomly scattering the mutations across residues covered in the structure (10,000,000 times). Both native (human) and homologous (>20% amino acid sequence identity) protein 3D structures were used in this analysis. Protein structures containing mutations from fewer than 5 samples were not analyzed because the results from such structures may lack robustness. In addition, we also used CLUMPS to assess enrichment of mutations at protein-protein interaction interfaces; this algorithm counts the mutations at residues located at protein interfaces and compares the count to a null model created by random mutational scattering of mutations across the structures. Images showing protein structures were created with Pymol v1.8.0.5 (http://pymol.org). Mutation diagrams (lollipop figures) of mutations were generated using *Mutations Mapper v1.0.1* (http://www.cbioportal.org/mutation_mapper.jsp)[8,9].

Correlation between driver genes and *GISTIC2.0* peaks. To investigate whether driver genes were more likely to be mutated in copy number regions or not, non-silent coding genes in the cohort were categorized by whether they were in a GISTIC peak that was affected by a copy number alteration in its patient, and whether it was in one of the driver genes identified as significantly mutated by *CLUMPS* or *MutSig2CV*. Fisher's exact test was then used to determine if mutations in driver genes co-occurred with significant copy number alterations more frequently than would be expected by random chance.

Assessment of chromothripsis

Due to the lack of full genome sequencing, we were limited to only one method of detection of chromothripsis as described in the literature[26]. To this end, each chromosome found to have been split into at least 10 segments longer than 100 exons in which there is at least one deletion of log2(CN/2)<0.25 and at least one gain of log2(CN/2)>0.25 and a variance at least 0.25 was tested to determine if the distances between the breakpoints were consistent with an exponential distribution, because it has been reported[33] that chromothripsis tends to deviate from this model while other mechanisms of copy number alterations do not.

Mutational Signature Analysis

Methods and Algorithms. The mutational signatures discovery is a process of de-convoluting cancer somatic mutations counts, stratified by mutation contexts or biologically meaningful subgroups, into a set of characteristic patterns (signatures) and inferring the activity of each of the discovered signatures across samples[27]. For this purpose, we exploited a Bayesian variant of non-negative matrix factorization (Bayesian

NMF) recently implemented and applied to several cancer genome projects (see [28,29] for additional background and technical details regarding the Bayesian NMF methodology). Bayesian NMF exploits a *shrinkage* or *automatic relevance determination* (ARD) technique to allow a sparse representation for both signatures and activities as well as an optimal inference for the number of signatures (*K*) by iteratively pruning away irrelevant components in balancing between a data-fidelity and a complexity[30]. The same parameters set as previously described were used[28,29]. All SNVs were classified to 96 possible mutation types or categories based on six base substitutions (C>A, C>G, C>T, T>A, T>C, and T>G) within the tri-nucleotide sequence context including the base immediately 5' and 3' to the mutated base.

Signature discovery in DLBCL 304 WES samples and identification of a micro-satellite unstable tumor. A de-novo signature extraction for 304 DLBCL WES samples with the BayesNMF applied to SNVs stratified by 96 tri-nucleotide mutation contexts identified three major mutational processes (**Supplementary Fig. 7a**). The similarity of these signatures to known 30 COSMIC signatures at http://cancer.sanger.ac.uk/cosmic/signatures was computed with a cosine similarity. The first signature (S304A; first *de novo* signature detected in 304 samples) characterized by a predominance of C>T mutations at CpG sites with minor contributions in C>A and T>C mutations was most similar to COSMIC6 (cosine similarity 0.87) and was exclusive to one sample with the highest mutation burden. The second signature (S304B; second *de novo* signature detected in 304 samples) characterized by a superposition of elevated C>T mutations at CpG sites with a background broad spectrum of base substitutions was most similar to the COSMIC1 (cosine similarity 0.86) and pervasive across samples, explaining about 64% overall mutations. The third signature (S304C; third *de novo* signature detected in 304 samples) was characterized by dominant T>G mutations at [C/G/T]pTpT sites with explaining about 7% overall mutations, but did not match 30 COSMIC signatures with the cosine similarity >= 0.78. Interestingly, the highest mutation burden sample (DLBCL-MAYO_DLBCL_234-Tumor; 5956 mutation) had a significantly higher activity of S304A (97% SNVs were associated with S304A). Since the signature profile of S304A most resembled the COSMIC6, which is known to be associated with defective DNA mismatch repairs and found in microsatellite unstable tumors, we further explored if this tumor had additional characteristics of the micro-satellite instability (MSI). We found this sample had a pathogenic splice site mutation in *MLH1* (chr3:37083822G>A) with a significant enrichment of insertions and deletions (14% in all variants). We also noted that the third highest mutation sample (DLBCL-RICOVER_787-Tumor-SM-4MILK) also had a relatively high activity of S304A (53% SNVs), and this sample had a pathogenic nonsense mutation in *MSH6* (R298*) with no indel enrichment.

Signature discovery in DLBCL 303 WES samples and identification of activation-induced cytidine deaminase signatures. To minimize a possible interference between the MSI signature (S304A) and the aging signature (S304B), which shared similar hotspot motifs, C>T at CpG sites, we excluded the putative MSI sample with the highest mutation burden in all downstream analyses and re-generate a de-novo signature extraction for 303 WES samples. In addition to 96 tri-nucleotide mutation types, we also considered the clustering information of mutations as an additional feature to capture a signal of the mutational process related to the activation-induced cytidine deaminase (AID signature). As was previously demonstrated[29], there was a substantial difference in mutation spectra between clustered and non-clustered mutations due to a differential activity of both canonical and non-canonical AID signatures. We first computed NMDs (Nearest Mutation Distance) for all SNVs, a minimum genomic distance to all other mutations on the same chromosome in the same patient, and partitioned them into 'clustered' (NMD <= 1kb) and 'nonclustered' groups (NMD > 1kb) (**Supplemental Fig.7b**). The threshold (1kb) was manually chosen from a bimodal feature of the NMD distribution. Then, we separately counted clustered and non-clustered mutations across 96 mutation channels and split mutations in each sample into two columns representing clustered and non-clustered mutational groups, giving rise to the mutation count matrix $\boldsymbol{X}$ (96 by 2*M*, *M* is the number of

samples). This mutation count matrix was ingested as an input for the BayesNMF and factored into two matrices, ***W'*** (96 by *K*) and ***H'*** (*K* by 2*M*), approximating ***X*** by ***W'H'***. It should be noted that clustered and non-clustered mutations from the same patient were separately handled to capture a characteristic signal from clustered mutations. Through a scaling transformation, $\boldsymbol{X} \sim \boldsymbol{W'H'} = \boldsymbol{WH}$, $\boldsymbol{W} = \boldsymbol{W'}\boldsymbol{U}^{-1}$ and $\boldsymbol{H} = \boldsymbol{UH'}$ where ***U*** is a *K*-by-*K* diagonal matrix with the element corresponding to the 1-norm of column vectors of ***W'***, resulted in the final signature loading matrix ***W*** and the activity loading matrix ***H***.

All fifty independent BayesNMF runs converged to the three signatures solution, identifying the aging signature (Aging), the canonical AID signature (cAID), and the secondary AID signature (AID2) shown in ***Fig. 2a***. The overall activity of discovered signatures in **Fig.2b** was determined by summing up the activities of three signatures assigned to both clustered (red) and non-clustered mutations (blue). The aging signature was characterized by pronounced C>T mutations at CpG sites superimposed with a background broad base substitutions, most similar to COSMIC1 (cosine similarity 0.93), and its activity was mostly attributed to non-clustered mutations (98% in non-clustered vs 2% in clustered aging mutations), explaining overall 80% SNVs across samples. The cAID signature had characteristic peaks of C>T and C>G mutations at GCT context corresponding to one of AID known hotspot motifs at RCY (R = A/G, Y= C/T), and its activity was much higher in clustered mutations (70% in clustered mutations) consistent to known AID biology. About 47% mutations related to the cAID signature was C>T or C>G mutations at RCY motifs. Interestingly, the third signature characterized by T>G mutations at [C/G]TT contexts also showed an enrichment of its activity in clustered mutations (36% in clustered mutations), and the signature profile was most similar to COSMIC9 (cosine similarity 0.75) corresponding to the non-canonical AID activity related to the error-prone DNA polymerase eta. Indeed, 50% mutations associated with the AID2 signature were A>[T/C/G] at WA (W=A/T) motifs corresponding to the hotspot motifs of non-canonical AID.

Assessment of the impact of a germline component in the mutational signature discovery. To address the impact of the germline contents in our tumor-only pipeline on the mutational signature discovery, we separately performed a signature discovery for the 134 samples with available patient-matched paired normals using either the tumor-normal pipeline (PAIRED-SET-TN, **Supplementary Fig. 7h**) or the tumor-only pipeline (PAIRED-SET-TO, **Supplementary Fig. 7i**). In both PAIRED-SET-TN and PAIRED-SET-TO, *SignatureAnalyzer* discovered three similar signatures highly concordant to those discovered in the COMBINED-SET (n=303, **Supplementary Fig. 7g**) irrespective of whether the samples were analyzed with their respective patient-matched normal samples or with our tumor-only pipeline (cosine similarity in **Supplementary Table 4f** and **Supplementary Fig. S7j**). Given that the germline component in tumor-only samples did not negatively impact the discovery of mutational signatures, we next evaluated if it skews the gene-level signature fractions for the significantly mutated genes (SMGs). To remove an additional confounding factor from the signature differences between PAIRED-SET-TN and PAIRED-SET-TO, we applied a projection approach to infer the signature activity of PAIRED-SET-TO samples onto the signature profiles of PAIRED-SET-TN. More specifically, the projection was done by minimizing the Kulbeck-Leibler divergence while the signature-loading matrix, ***W***, comprised of the column vectors corresponded to normalized signature profiles of PAIRED-SET-TN (Aging, cAID, and AID2) is frozen, and the activity- loading matrix **H** is iteratively updated to best approximate the mutation count matrix of PAIRED-SET-TO, **X.** The resulting row vectors in **H** represent a de-convoluted signature activity of PAIRED-SET-TO samples onto the signatures of PAIRED-SET-TN. We found a strong correlation of the paired signature fraction for all SMGs for the aging, cAID signature and AID2 signature (Pearson correlation = 0.96, 0.98 and 0.95, respectively; **Supplementary Fig. 7k-m**).

Assesement of the impact of FFPE artifacts in the mutational signature discovery. We also performed additional analyses to assess the impact of FFPE bias on mutational signature discovery. We separately applied *SignatureAnalyzer* to tumors from frozen (FRZN-SET, **Supplementary Fig. 7n**) and FFPE tissue (FFPE-SET, **Supplementary Fig. 7o**) and compared discovered signatures to those in COMBINED-SET (**Supplementary Fig. 7p** and **Supplementary Table 4g**). We observed that the cosine similarity of both C>T_CpG and cAID signatures was very high among the three sets (**Supplementary Fig. 7p** and **Supplementary Table 4g**), while the AID2 signature in FFPE-SET has a slightly reduced similarity. To investigate the effects of FFPE samples more systematically we performed a series of signature discoveries for the pooled sample sets generated by randomly replacing fresh-frozen samples by the same number of random FFPE samples (500 experiments). We observed strong stability in the cosine similarities of both aging and cAID signatures, but a subtle drop in the AID2 signature with incremental fractions of FFPE samples (**Supplementary Fig. 7q**). However, we cannot rule out that this reduced cosine similarity of the AID2 signature might be attributed to the sample heterogeneity between FRZN-SET and FFPE-SET.

Signature enrichment analysis. We first annotated each mutation with the probability (likelihood of association) that it was generated by each of the discovered mutational signatures, P_{ms}, where '*m*' denoted a mutation and '*s*' refers to the signature. More specifically, the likelihood of association to the *k*-th signature for a set of mutations corresponding to *i*-th mutation context and *j*-th clustered or non-clustered mutation group was defined as $[\boldsymbol{w}_k\boldsymbol{h}_k / \sum(\boldsymbol{w}_k\boldsymbol{h}_k)]_{ij}$, where $\boldsymbol{w}_k$ and $\boldsymbol{h}_k$ correspond to the *k*-th column vector and *k*-th row vector of ***W*** and ***H***, respectively. The relative activity enrichment for candidate cancer genes (CCGs) with at least 10 mutations in ***Fig. 2c*** was determined by taking an average of P_{ms} for all mutations in each CCGs. For the gene-level signature-enrichment analysis, we first attempted to identify a hotspot mutation motif out of 96 contexts in each signature by considering coding mutations only with P_{ms} >= 0.75, identifying 40 and 50 characteristic motifs with non-zero probability for cAID and AID2, respectively. Note that keeping mutations with a higher P_{ms}, filtered out mutations shared by multiple signatures and enabled the discovery of more distinct mutation motifs characteristic to each signature. To take into account sequence composition variation across the genome, we enumerated all available tri-nucleotide contexts across coding genes and considered genes having non-zero mutations with P_{ms} >= 0.75 in each signature. This information was used to estimate the background mutation rates at the hotspot motifs in each signature, resulting in r_{Aging} = 4.3 per Mb, r_{cAID} = 7.5 per Mb and r_{AID2} = 2.4 per Mb for the aging, cAID, and AID2 signatures, respectively. Then, for given mutation counts with P_{ms} >= 0.75, *x*, at hotspot motifs and available sequence context, *n*, in each gene, we performed a one-sided binomial test with the estimated background mutation rate to assess the significance of the enrichment of each signature across 12532 genes for the aging, 328 genes for cAID, and 967 genes for AID2 signature having non-zero mutations with P_{ms} >= 0.75 (**Supplementary Table 4b-d**). We corrected for multiple hypotheses and identified genes that are associated with each signature using a *q*-value cutoff of 0.1 (see Q–Q plots in **Supplementary Fig. 7c**).

Statistical analysis related to signatures. The age correlation with three signatures in **Supplementary Fig. 7d** was performed by binning the age into seven groups and calculating a Pearson correlation between the median age and the median activity in each age group. The absolute signature activity in ***Fig. S14e*** (left) was computed by first binning the CCF of all SNVs into ten groups and counting the number of mutation with P_{ms} >= 0.5 in each CCF bin, and the relative signature activity (**Supplementary Fig. 14e**, right) was defined as a fraction of the absolute signature activity to the total number of SNVs in each CCF bin. To determine if the AID mutations discovered in this cohort had clustering around the transcription start site (TSS)[31], a Fisher's exact test was applied to determine if AID mutations (P_{cAID} + P_{AID2} > 0.75) in our 98 CCG were more frequently within +/- 2000 base pairs of its TSS. We used a Wilcoxon rank sum test to determine if mutations

from the 98 CCGs attributed to either of the two AID signature had a significantly higher proportion of silent mutations than mutations not attributed to either AID signature.

Integrative analysis of gene expression and copy number data.

Gene expression profiling and data normalization. RNA samples from 52 samples with available WES data were transcriptionally profiled using an U133plus2 Affymetrix gene expression array as previously described ($batch_1$)[4]. The data has been uploaded to GEO with the accession GSE98588. Additional expression profiles from 85 samples with available WES data were generated and published previously (GSE34171, $batch_2$)[4]. The Affymetrix gene expression profiles were normalized using Robust Multi-Array Average (RMA)[32] and Brainarray custom chip definition files (Version 16) based on Ensemble IDs[33]. Gene expression values were adjusted for batch effect using a linear regression model of each gene against the batch variable ($batch_1$ vs. $batch_2$). The batch-corrected gene expression values are the residuals of the linear regression plus the intercept. Log_2–transformed batch-corrected gene expression was used for differential analysis. The following analyses were performed using the integrative (Epi)DNA-to-Gene Expression analysis package (iEDGE, manuscript in preparation).

Cis-analysis. Genomic coordinates of genes within SCNAs were determined using R Bioconductor annotation package *TxDb.Hsapiens.UCSC.hg19.knownGene*. Genes within arm-level alterations were considered for within-arm differential expression analysis (**Supplementary Table 6a**). Genomic boundaries of chromosome arms were defined by the start, centromere, and end coordinates of each chromosome as annotated in the UCSC hg19 cytoband annotation file. Separately, genes with coordinates within wide peak limits of each GISTIC2-defined copy number alteration with a FDR q-value <0.1 were considered for within-peak differential expression analysis (**Supplementary Table 6a**).

Expression of the genes within each GISTIC-defined alteration peak was tested for association with the corresponding peak's somatic copy number alteration (SCNA) status (presence or absence of copy gain or loss) by a differential expression test using the R package limma[34]. One-sided p-values were estimated, since the associations of interest are gene expression up-regulation among samples with copy gain and gene expression down-regulation among samples with copy loss. The p-values for all genes across all alteration peaks were corrected for multiple hypothesis testing using the false discovery rate (FDR) estimation[35]. Genes with FDR < 0.25 and a fold change of >1.2 were considered significant "cis-acting" genes. We performed 3 types of differential expression analysis: 1. Arm-levels: Integrating arm-level SCNA status and expression of genes on each arm (**Supplementary Table 6b-c**); 2. Focal alterations: Integrating focal SCNAs and expression of genes in focal peaks (**Supplementary Table 6d-e**), and 3. focal or arm: comparing focal or arm level SCNA status and expression of genes in focal peak. In this case, the copy number alteration status of the sample is considered altered if it is altered in the focal peak or the arm that harbors the focal peak. (**Supplementary Table 6f-g**).

For focal events, COSMIC cancer genes[36] with a positive correlation to gene expression in our data (fold change >1.2, q<0.25) are indicated in **Fig 4a**.

Cell-of-origin (COO) assignment. The COO phenotype was assigned for 80% (242/304) of samples. For the newly gene expression profiled fresh frozen samples (GSE98588, $batch_1$), the COO assignment was performed using a linear-predictive-score classifier as previously published (**Supplementary Table 1**)[4,37,38]. The COO phenotypes for the 85 previously published samples (GSE34171, $batch_2$) were previously reported

(***Table S1***)[4]. NanoString-based COO assignment using the Lymph2Cx assay[39] was performed for additional 102 FFPE samples as recently reported (**Supplementary Table 1**)[40].

Gene Set Enrichment Analysis (GSEA). For samples with paired available gene expression data, GSEA was performed as previously described[4,21,41]. Indicated target gene sets (**Supplementary Fig. 13f-h**) were tested for an enrichment in a given DLBCL clusters vs. the union of samples in the other DLBCL clusters (excluding C0 DLBCLs).

Targeted DNA-sequencing for the detection of chromosomal rearrangements

Library Construction, sequencing and pre-analysis processing. Targeted rearrangements (Supplementary Table 5a) were captured from either leftover uncaptured libraries from WES or genomic DNA, sequenced using an Illumina sequencing platform, de-multiplexed and aligned to the reference sequence b37 edition from the Human Genome Reference Consortium with bwa[42] as described previously[43,44]. A total of 296/304 samples had a mean read depth is 221.4x and met all quality control checkpoints and 99% of samples had a power greater than 0.996 to detect chromosomal rearrangements.

Chromosomal rearrangement pipeline. Somatic rearrangements were detected using four different calling algorithms, *BreaKmer[18], Lumpy[17], dRanger[16]* and *SVaBA[45],* followed by *Breakpointer* validation, filtering and a CCF estimation module (**Supplementary Fig. 8a**) as described below.

Chromosomal rearrangement detection. BreaKmer[18], Lumpy[17] and dRanger[16] were applied as previously described to generate a separate list of candidate rearrangements.

SVaBA. SVaBA identified rearrangements by performing *de novo* local assembly across every 25 kb region in the genome, with 1 kb overlaps. Assembly within *SVaBA* was achieved through a modified version of SGA[46], which assembled reads with gapped alignments, unmapped pair mates, clipped alignments and reads with an aligned insert size that differed substantially from the mean. The assembled contigs were re-aligned within *SVaBA* to the reference genome using an in-memory implementation of BWA-MEM[47,48]. Contigs with multi-part alignments were used to infer candidate rearrangements, excluding contigs with low alignment quality. Within each local assembly window, rearrangements were genotyped by finding the optimal alignment of the sequencing reads to either the variant-supporting contig or the reference genome. Rearrangements obtained from contigs with breakpoints supported by > 4 tumor reads and no normal reads were classified as somatic. In addition to detecting rearrangements from assembled contigs, discordant reads were clustered as a second signal for rearrangements. In the absence of a supporting contig, discordant read clusters required a minimum of 8 tumor read pairs with a high alignment quality, and no normal read pairs, to be called as a somatic rearrangement. The discordant read clusters were further compared with the rearrangements obtained from the contig assemblies to obtain the total number of variant supporting reads for each somatic rearrangement.

Filtering and Breakpointer module. The candidate rearrangements detected by each detector were filtered for variants found in 342 in house normal samples from the Broad Institute based on a 5 kb window to match candidate rearrangements. The remaining variants were clustered based on a 50 bp window and the union of all unique clustered candidate rearrangements was passed to *Breakpointer[16]* (**Supplementary Fig. 8a**). *Breakpointer* scans for additional supporting split read evidence to confirm breakpoint junctions in the tumor sample and to reject candidate somatic rearrangements with evidence of the rearrangement appearing in matched normal samples. We required a combined total of at least 4 supporting reads, either read pairs or split reads, following *Breakpointer*. We also used a panel of an additional 21 normal samples sequenced

with the same targeted bait set and protocol to reject artifacts specific to the targeted protocol. A 5 kb window was used to match candidate rearrangements. In addition, we filtered out all intrachromosomal rearrangements (deletions and tandem fusions) involving the *IGH, IGK, IGL, TRA, TRB, TRG* loci and translocations between two loci, both not part of the bait set.

A total of 3293 structural variants were called and passed all standard panel of normal filtering steps (1355 from dRanger, 514 from *SVaBA*, 453 from Breakmer and 1775 from Lumpy). Subsequent clustering, *Breakpointer* validation, filtering events found in 19 similar processed normal samples and post-processing review resulted in 413 reported SVs (**Supplementary Table 5**).

CCF calculation for SVs. For SVs, we applied a novel algorithm for determining CCF (C_{SV}) based on local copy number and allele fraction. The calculation here is roughly equivalent of the *ABSOLUTE* recipe of mutation CCFs, except that SVs consist of two breakpoints, each with its own estimated allele fraction, underlying copy number, and multiplicity. SV multiplicity has the same meaning as mutation multiplicity: the number of SV events per cell. With targeted data, there is an additional complication in that not all the breakpoints occur within targeted regions, which leaves the observed allele counts biased against the reference allele. For this reason, only breakpoints found within a targeted region are used in SV CCF estimates. At a given breakpoint, the DNA copy state is defined according to:

$$D = (T_a + T_b) * C_{CNV} * \alpha + (N_a + N_b) * (1 - C_{CNV} * \alpha)$$

Where

$\alpha = \text{tumor purity}$
$C_{CNV} = \text{Cancer Cell Fraction of cells with SCNA state}$
$T_a = \text{minor copy number allele in tumor cells}$
$T_b = \text{major copy number allele in tumor cells}$
$N_a = \text{minor copy number allele in normal cells}$
$N_b = \text{major copy number allele in normal cells}$

SV breakpoints often occur at edges of copy number segments, which introduces some ambiguity regarding the relevant copy number state. The copy number estimates used here are those within the SV alternate allele, which corresponds to a window upstream (3' direction in reference coordinates) of forward mapped alt supporting reads and downstream (5') of reverse mapped alt supporting reads for each breakpoint. We used a 10kb window, roughly consistent with the breakpoint resolution of our copy number segmentation algorithm (*ReCapSeg*). Estimates of T_a, T_b, and C_{CNV} are based on *ABSOLUTE* and *AllelicCapseg*. Although the copy number state and the SV allele fractions may not be the same at both breakpoints for a given SV, the SV CCF is constrained to be the same at both breaks. At each break, the CCF is estimated as a CCF probability density distribution (pdf) and the combined SV CCF is the joint pdf from the two breakpoint CCF pdfs. In cases where only one of the two breakpoints was contained within a targeted region, only the targeted breakpoint CCF pdf was used to estimate the SV CCF. The expected allele fraction for a given SV breakpoint is calculated for different somatic variant configurations similar to the different scenarios modeled when calculating the germline/somatic for point mutations. There are three basic SV event scenarios that depend on the ordering of events (SV comes before or after the SCNA) and whether or not the SV occurs in tumor cells with or without the somatic copy number variant:

1) The SV variant occurs on cells with somatic copy number variants (T_a or T_b) with multiplicity m. In this scenario the SV event occurs chronologically after the copy number event so the multiplicity(*m*) should be 1.

$$AF_1 = \frac{\alpha m C_{SV}}{D}$$

2) The SV variant occurs in tumor cells without the somatic copy number alteration with multiplicity 1.

$$AF_2 = \frac{\alpha C_{SV}}{D}$$

3) The SV variant occurs in all cells with the the somatic copy number alteration Ta or Tb with multiplicity m and a fraction of cells without the somatic copy number alteration. In this scenario the SV event occurs chronologically before the copy number event.

$$AF_3 = \frac{\alpha(C_{SV} + (m-1) * C_{CNV})}{D}$$

C_{SV}, the CCF of the structural variant, from f 0 to 1 in increments of 0.01 to construct the C_{SV} pdf. The C_{SV} pdf is based on the beta pdf of the estimated AF with the observed read depth d, and the counts of alternate allele supporting reads a.

$$pdf(a|d, C_{SV}) = \beta(AF_i, a+1, d-a+1) = \frac{{AF_i}^a (1-AF_i)^{d-a} * \Gamma(d+2)}{\Gamma(a+1)\Gamma(d-a+1)}$$

where β is the beta probability density distribution for observing a alt reads, from a total of d with allele fraction AF_i. The scenario with the maximum pdf mode was chosen to represent the SV breakpoint. This boils down to a choice between scenario 3 and 1 since scenario 1 and 2 are mathematically equivalent when the SV multiplicity is 1. The combined SV CCF from both breakpoints is the joint pdf:

$$pdf(a_1, a_2|d_1, d_2, C_{SV}) = pdf(a_1|d_1, C_{SV}) * pdf(a_2|d_1, d_2, C_{SV})$$

The C_{SV} value for a given SV is the mode of the CCF pdf distribution and the 95% confidence interval for C_{SV} is the 95% region of the normalized pdf around the mode.

Validation of CCF in LBCL cell lines. The CCF calculation for structural variants was also applied to 31 B-cell lymphoma (DLBCL and follicular lymphoma) cell lines as a validation test of the method, with the assumption that the bulk of known driver SV events in these cell lines should be clonal. Sequencing data for the 31 cell lines used the same protocol with as the targeted data for the detection of SVs in DLBCL. The results of this test are shown in **Supplementary Fig. 14h**, which show that the bulk of driver translocations *IgH-BCL2, IgH-BCL6* and *IgH-MYC* are found to be clonal with CCFs exceeding 0.9. Only one cell line, Ly18, had driver SVs but neither of the balanced translocations between *MYC* and *IGH* had CCF 95% CI's that excluded CCF >0.9. Eight cell lines lacked a driver SV and the remaining 22 cell lines had at least one clonal driver SV.

Visualization. Rearrangements were visualized either as circos plots (http://circos.ca) or as stick figures plotting the breakpoint in its genomic context.

Immunohistochemistry of PD-1 ligands

Double staining of PD-L1 (Cell Signaling, clone 405.9A11) and PAX5 (BD Biosciences, 24/Pax-5) and staining of PD-L2 (EMD Milipore, clone 366C.9E5) was performed with an automated staining system (Bond III; Leica Biosystems, Buffalo Grove) as previously described[43,49].

Supplementary Bibliography

1. Pfreundschuh, M.*, et al.* Six versus eight cycles of bi-weekly CHOP-14 with or without rituximab in elderly patients with aggressive CD20+ B-cell lymphomas: a randomised controlled trial (RICOVER-60). *Lancet Oncol* **9**, 105-116 (2008).
2. Lohr, J.G.*, et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc Natl Acad Sci U S A* **109**, 3879-3884 (2012).
3. Novak, A.J.*, et al.* Whole-exome analysis reveals novel somatic genomic alterations associated with outcome in immunochemotherapy-treated diffuse large B-cell lymphoma. *Blood Cancer Journal* **5**, e346 (2015).
4. Monti, S.*, et al.* Integrative Analysis Reveals an Outcome-Associated and Targetable Pattern of p53 and Cell Cycle Deregulation in Diffuse Large B Cell Lymphoma. *Cancer Cell* **22**, 359-372 (2012).
5. Lichtenstein, L., Wood, B., MacBeth, A., Birsoy, O. & Lennon, N. Abstract 3641: ReCapSeg: Validation of somatic copy number alterations for CLIA whole exome sequencing. *Cancer Research* **76**, (14 Supplement) 3641; DOI: 3610.1158/1538-7445.AM2016-3641 (2016).
6. Lawrence, M.S.*, et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218 (2013).
7. Reddy, A.*, et al.* Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell* **171**, 481-494 e415 (2017).
8. Cerami, E.*, et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* **2**, 401-404 (2012).
9. Gao, J.*, et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1 (2013).
10. Wan, P.T.*, et al.* Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell* **116**, 855-867 (2004).
11. Dhillon, A.S. & Kolch, W. Oncogenic B-Raf mutations: crystal clear at last. *Cancer Cell* **5**, 303-304 (2004).
12. Holderfield, M.*, et al.* RAF inhibitors activate the MAPK pathway by relieving inhibitory autophosphorylation. *Cancer Cell* **23**, 594-602 (2013).
13. Kamburov, A.*, et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc Natl Acad Sci U S A* **112**, E5486-5495 (2015).
14. Betts, M.J.*, et al.* Mechismo: predicting the mechanistic impact of mutations and modifications on molecular interactions. *Nucleic Acids Res* **43**, e10 (2015).
15. Welcker, M. & Clurman, B.E. FBW7 ubiquitin ligase: a tumour suppressor at the crossroads of cell division, growth and differentiation. *Nat Rev Cancer* **8**, 83-93 (2008).
16. Drier, Y.*, et al.* Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res* **23**, 228-235 (2013).
17. Layer, R.M., Chiang, C., Quinlan, A.R. & Hall, I.M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**, R84 (2014).
18. Abo, R.P.*, et al.* BreaKmer: detection of structural variation in targeted massively parallel sequencing data using kmers. *Nucleic Acids Res* **43**, e19 (2015).
19. Valls, E.*, et al.* BCL6 Antagonizes NOTCH2 to Maintain Survival of Human Follicular Lymphoma Cells. *Cancer Discovery* **7**, 506-521 (2017).
20. Beguelin, W.*, et al.* EZH2 is required for germinal center formation and somatic EZH2 mutations promote lymphoid transformation. *Cancer Cell* **23**, 677-692 (2013).
21. Subramanian, A.*, et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
22. Landau, D.A.*, et al.* Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525-530 (2015).
23. Cibulskis, K.*, et al.* ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601-2602 (2011).
24. Taylor-Weiner, A.*, et al.* DeTiN : Overcoming Tumor in Normal Contamination. *Nature Methods* (2018) *accepted*.

25. Berman, H.M.*, et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).
26. Korbel, J.O. & Campbell, P.J. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152**, 1226-1236 (2013).
27. Alexandrov, L.B.*, et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402-1407 (2015).
28. Kim, J.*, et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet* **48**, 600-606 (2016).
29. Kasar, S.*, et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat Commun* **6**, 8866 (2015).
30. Tan, V.Y. & Fevotte, C. Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. *IEEE Trans Pattern Anal Mach Intell* **35**, 1592-1605 (2013).
31. Di Noia, J.M. & Neuberger, M.S. Molecular mechanisms of antibody somatic hypermutation. *Annu Rev Biochem* **76**, 1-22 (2007).
32. Irizarry, R.A.*, et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264 (2003).
33. Dai, M.*, et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* **33**, e175 (2005).
34. Ritchie, M.E.*, et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
35. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).
36. Futreal, P.A.*, et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183 (2004).
37. Monti, S.*, et al.* Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood* **105**, 1851-1861 (2005).
38. Wright, G.*, et al.* A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proceedings of the National Academy of Sciences* **100**, 9991-9996 (2003).
39. Scott, D.W.*, et al.* Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood* **123**, 1214-1217 (2014).
40. Staiger, A.M.*, et al.* Clinical Impact of the Cell-of-Origin Classification and the MYC/ BCL2 Dual Expresser Status in Diffuse Large B-Cell Lymphoma Treated Within Prospective Clinical Trials of the German High-Grade Non-Hodgkin's Lymphoma Study Group. *J Clin Oncol* **35**, 2515-2526 (2017).
41. Chapuy, B.*, et al.* Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. *Cancer Cell* **24**, 777-790 (2013).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
43. Chapuy, B.*, et al.* Targetable genetic features of primary testicular and primary central nervous system lymphomas. *Blood* **127**, 869-881 (2016).
44. Chapuy, B.*, et al.* Diffuse large B-cell lymphoma patient-derived xenograft models capture the molecular and biological heterogeneity of the disease. *Blood* **127**, 2203-2213 (2016).
45. Wala, J.*, et al.* Genome-wide detection of structural variants and indels by local assembly. *bioRxiv* (2017).
46. Simpson, J.T. & Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* **22**, 549-556 (2012).
47. Wala, J. & Beroukhim, R. SeqLib: a C ++ API for rapid BAM manipulation, sequence alignment and sequence assembly. *Bioinformatics* (2016).
48. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997* (2015).
49. Roemer, M.G.*, et al.* PD-L1 and PD-L2 Genetic Alterations Define Classical Hodgkin Lymphoma and Predict Outcome. *J Clin Oncol* **34**, 2690-2697 (2016).